



# **Four Essays in Quantitative Analysis: Artificial Intelligence and Statistical Inference**

Arman Hassanniakalager

A Dissertation

Submitting in fulfilment of the requirements of the  
Degree of Doctor of Philosophy

Adam Smith Business School  
College of Social Sciences  
University of Glasgow

August 2018

## Abstract

This thesis consists of four essays exploring quantitative methods for investment analysis. Chapter 1 is an introduction to the topic where the backgrounds, motivations and contributions of the thesis are discussed. This Chapter proposes an expert system paradigm which accommodates the methodology for all four empirical studies presented in Chapters 2 to 5.

In Chapter 2 the profitability of technical analysis and Bayesian Statistics in trading the EUR/USD, GBP/USD, and USD/JPY exchange rates are examined. For this purpose, seven thousand eight hundred forty-six technical rules are generated, and their profitability is assessed through a novel data snooping procedure. Then, the most promising rules are combined with a Naïve Bayes (NB), a Relevance Vector Machine (RVM), a Dynamic Model Averaging (DMA), a Dynamic Model Selection (DMS) and a Bayesian regularised Neural Network (BNN) model. The findings show that technical analysis has value in Foreign eXchange (FX) trading, but the profit margins are small. On the other hand, Bayesian Statistics seems to increase the profitability of technical rules up to four times.

Chapter 3 introduces the concept of Conditional Fuzzy (CF) inference. The proposed approach is able to deduct Fuzzy Rules (FRs) conditional to a set of restrictions. This conditional rule selection discards weak rules and the generated forecasts are based only on the most powerful ones. In order to achieve this, an RVM is used to extract the most relevant subset of predictors as the CF inputs. Through this process, it is capable of achieving higher forecasting performance and improving the interpretability of the underlying system. The CF concept is applied in a betting application on football games of three main European championships. CF's performance in terms of accuracy and profitability over the In-Sample (IS) and Out-Of-Sample (OOS) are benchmarked against the single RVM and an Adaptive Neuro-Fuzzy Inference System (ANFIS) fed with the same CF inputs and an Ordered Probit (OP) fed with the full set of predictors. The results demonstrate that the CF is providing higher statistical accuracy than its benchmarks, while it is offering substantial profits in the designed betting simulation.

Chapter 4 proposes the Discrete False Discovery Rate (DFDR<sup>+/·</sup>) as an approach to compare a large number of hypotheses at the same time. The presented method limits the probability of having lucky findings and accounts for the dependence between candidate models. The performance of this approach is assessed by backtesting the predictive power of technical analysis in stock markets. A pool of twenty-one thousand technical rules is tested for a positive Sharpe ratio. The surviving technical rules are used to construct dynamic portfolios. Twelve categorical and country-specific Morgan Stanley Capital International (MSCI) indexes are examined over ten years (2006-2015). There are three main findings. First, the proposed method has high power in detecting the profitable trading strategies and the time-related anomalies across the chosen financial markets. Second, the emerging and frontier markets are more profitable than the developed markets despite having higher transaction costs. Finally, for a successful portfolio management, it is vital to rebalance the portfolios on a monthly basis or more frequently.

Chapter 5 undertakes an extensive investigation of volatility models for six securities in FX, stock index and commodity markets, using daily one-step-ahead forecasts over five years. A discrete false discovery controlling procedure is employed to study one thousand five hundred and twelve volatility models from twenty classes of Generalized AutoRegressive Conditional Heteroskedasticity (GARCH), Exponential Weighted Moving Average (EWMA), Stochastic Volatility (SV), and Heterogeneous AutoRegressive (HAR) families. The results indicate significant differences in forecasting conditional variance. The most accurate models vary across the three market categories and depend on the study period and measurement scale. Time-varying means, Integrated GARCH (IGARCH) and SV, as well as fat-tailed innovation distributions are the dominant specifications for the outperforming models compared to three benchmarks of ARCH (1), GARCH (1,1), and the volatility pool's 90<sup>th</sup> percentile.

Finally, Chapter 6 puts together the main findings from the four essays and presents the concluding marks.

# Table of Contents

Abstract .....	ii
Table of Contents .....	iv
List of Tables.....	viii
List of Figures .....	xii
Acknowledgements.....	xiv
Author's Declaration .....	xvi
Abbreviations .....	xvii
1. Introduction.....	1
1.1 Background and Motivation .....	1
1.2 Contribution and Structure .....	3
2. Trading the Foreign Exchange Market with Technical Analysis and Bayesian Statistics .....	6
2.1 Introduction .....	6
2.2 Methodology.....	11
2.2.1 Data Snooping Test.....	11
2.2.2 RVM.....	13
2.2.3 DMA and DMS .....	17
2.2.4 BNN .....	20
2.2.5 NB.....	21
2.3 Empirical Section .....	23
2.3.1 Dataset.....	23
2.3.2 Trading Application .....	24
2.3.3 Bayesian Methods.....	25
2.4 Conclusions.....	27
3. Conditional Fuzzy Inference: Applications in Football Results Forecasting ...	28
3.1 Introduction .....	28
3.2 Methodology.....	34

3.2.1 RVM - Underlying System .....	34
3.2.2 Conditional Fuzzy Inference .....	34
3.2.2.1 Motives .....	34
3.2.2.2 Algorithm.....	35
3.2.3 Benchmarks.....	42
3.2.4 Kelly Criterion Application .....	43
3.3 Empirical Study .....	44
3.3.1 Dataset.....	45
3.3.2 Empirical Results .....	47
3.4 Conclusions.....	51
4. Technical Analysis and the Discrete False Discovery Rate: Evidence from MSCI Indexes.....	53
4.1 Introduction .....	53
4.2 Methodology.....	58
4.2.1 Overview of the FDR Procedure .....	58
4.2.2 Issues Regarding Existing FDR Methods.....	60
4.2.3 $DFDR^{+/-}$ .....	63
4.2.4 FDR Portfolio Construction.....	65
4.3 Dataset, Technical Trading Rules and Performance metrics.....	66
4.3.1 Dataset.....	66
4.3.2 Technical Rules .....	67
4.3.3 Excess Returns, Transaction Costs and Performance metrics.....	67
4.4 IS Performance .....	70
4.4.1 Identification and IS Profitability .....	71
4.4.2 Break-even Transaction Costs .....	71
4.5 OOS Performance.....	73
4.5.1 Excess Return .....	73
4.5.2 Performance Persistence .....	77
4.5.3 OOS Cross-validation.....	78

4.5.4 Financial Stress .....	81
4.6 Conclusion .....	83
5. Revisiting Financial Volatility Forecasting: Evidence from Discrete False Discovery Rate .....	85
5.1 Introduction .....	85
5.2 The Conditional Volatility Pool .....	88
5.2.1 Conditional Mean and Variance .....	89
5.2.2 Forecasting Models .....	90
5.3 DFDR <sup>+</sup> .....	92
5.4 Model setup .....	92
5.4.1 Data .....	92
5.4.2 Performance Metrics .....	93
5.5 Results .....	94
5.5.1 Model Performance.....	94
5.5.2 True Discoveries .....	96
5.5.3 Distribution .....	96
5.5.4 Mean Estimation .....	97
5.5.5 Conditional Variance .....	98
5.5.6 Class .....	99
5.6 Conclusion .....	100
6. Conclusion .....	102
6.1 Summary .....	102
6.2 Limitations .....	103
6.3 Future Works.....	104
Appendices .....	106
Appendix A .....	106
A.1 Technical Trading Rules .....	106
A.2 Sharpe Ratio .....	111
Appendix B .....	112

B.1 CF Illustration Example .....	112
B.2 IS Performance .....	114
B.3 CF Games .....	115
B.4 ANFIS .....	115
B.5 OP .....	117
Appendix C .....	119
C.1 Monte Carlo simulations .....	119
C.2 Robustness Exercise .....	123
Appendix D .....	128
D.1 Specification of the Pool .....	128
D.2 Model Selection Algorithm .....	128
D.3 Densities for Other loss Functions.....	129
D.4 True Discoveries Dynamics over Time for Other Distributions .....	132
D.5 Error Distribution Analysis for Other Loss Functions .....	132
D.6 Mean Analysis of Other Loss Functions .....	133
D.7 Conditional Variance Analysis for Other Loss Functions .....	135
D.8 Class Analysis for Other Loss Functions .....	138
Bibliography .....	139
Tables .....	161

## List of Tables

Table 2.1: Descriptive Statistics .....	161
Table 2.2: Trading Rules Excess Annualized Return and Sharpe Ratio .....	162
Table 2.3: EUR/USD Trading Performance - Annualized Return.....	163
Table 2.4: GBP/USD Trading Performance - Annualized Return.....	164
Table 2.5: USD/JPY Trading Performance - Annualized Return .....	165
Table 3.1: Football Forecasting Literature Comparison .....	166
Table 3.2: Inputs Series .....	167
Table 3.3: Accuracy Ratios (Game Result).....	168
Table 3.4: Accuracy Ratios (Asian Handicap).....	169
Table 3.5: Accuracy Ratios (Number of Goals).....	170
Table 3.6: Average Profit per Bet (Game Result).....	171
Table 3.7: Average Profit per Bet (Asian Handicap).....	172
Table 3.8: Average Profit per Bet (Number of Goals).....	173
Table 3.9: Proportional Cumulative Annualized Return (Game Result) .....	174
Table 3.10: Proportional Cumulative Annualized Return (Asian Handicap).....	175
Table 3.11: Proportional Cumulative Annualized Return (Number of Goals) ....	176
Table 3.12: Kelly Criterion (CF) .....	177
Table 3.13: Kelly Criterion (OP).....	178
Table 4.1: Summary statistics of the daily return series under study (12 MSCI indexes and the federal funds rate). .....	179
Table 4.2: Percentage and standard deviation of the DFDR <sup>+/-</sup> procedure survivors (IS 2 years).....	180
Table 4.3: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years) .....	181
Table 4.4: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 1 Month) .....	182
Table 4.5: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 3 Months) .....	183



Table 4.6: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 6 Months) .....	184
Table 4.7: Monthly Performance Persistence for IS 2 Years (1 month rolling-forward) .....	185
Table 4.8: Quarterly Performance Persistence in Months for IS 2 2 Years (3 months rolling-forward) .....	186
Table 4.9: Semi-annual Performance Persistence in Months for IS 2 Years (6 months rolling-forward) .....	187
Table 4.10: Annualized Returns Based on the Cross-validated Surviving Rules (IS of 2 Years and OOS 1 Month) .....	188
Table 4.11: Annualized Returns Based on the Cross-validated Surviving Rules (IS of 2 Years and OOS 3 Months) .....	189
Table 4.12 Annualized Returns Based on the cross-validated surviving rules (IS of 2 Years and OOS 6 Months).....	190
Table 4.13: Financial Stress Performance .....	191
Table 5.1: List of GARCH, SV, EWMA, and HAR volatility models. ....	192
Table 5.2: Summary Statistics of Log Returns.....	193
Table 5.3: Loss Functions .....	194
Table 5.4: Variation in Number of True Discoveries .....	195
Table 5.5: Innovation Distribution Survival Rate Across the Markets. ....	196
Table 5.6: Classes Survival Rates .....	197
Table A.1: EUR/USD Trading Performance - Sharpe Ratio .....	198
Table A.2: GBP/USD Trading Performance - Sharpe Ratio .....	199
Table A.3: USD/JPY Trading Performance - Sharpe Ratio .....	200
Table B.1: Relevance Vectors .....	201
Table B.2: Cluster Characteristics for the Generated Rules .....	202
Table B.3: Regression coefficients for the generated rules.....	203
Table B.4: IS Accuracy .....	204
Table B.5: CF Forecasts .....	205

Table C.1: Annualized Mean Excess Returns for Quartiles of Different Combination of Sharpe Ratio Levels. ....	206
Table C.2 Estimation of Neutral, Positive, and Negative Proportions by the DFDR <sup>+/-</sup> Procedure Versus the Actual Ones. ....	207
Table C.3: True FDR, Accuracy and the Positive-performing Portfolio Size through Different Methods. ....	208
Table C.4: Percentage and Standard Deviation of the DFDR <sup>+/-</sup> Procedure Survivors (IS 1 Year). ....	209
Table C.5: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year) ....	210
Table C.6 Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 1 Month) ....	211
Table C.7: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 3 Months) ....	212
Table C.8: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 6 Months) ....	213
Table D.1: Specification of the Volatility Forecasting Pool ....	214
Table D.2: Variation in Number of True Discoveries for MAE1 ....	215
Table D.3: Variation in Number of True Discoveries for MAE2 ....	216
Table D.4: Variation in Number of True Discoveries for MSE2 ....	217
Table D.5: Variation in Number of True Discoveries for R2LOG ....	218
Table D.6: Variation in Number of True Discoveries for QLIKE.....	219
Table D.7: Innovation Distribution for MAE1 ....	220
Table D.8: Innovation Distribution for MAE2 ....	221
Table D.9: Innovation Distribution for MSE2 ....	222
Table D.10: Innovation Distribution for R2LOG ....	223
Table D.11: Innovation Distribution for QLIKE ....	224
Table D.12: Classes Survival Analysis for MAE1 ....	225
Table D.13: Classes Survival Analysis for MAE2 ....	226
Table D.14: Classes Survival Rate for MSE2 ....	227

Table D.15: Classes Survival Rate for R2LOG.....228

Table D.16: Classes Survival Rate for QLIKE .....229

## List of Figures

Figure 1.1: Proposed Paradigm for Financial Decision Making with Examples.....	3
Figure 3.1: ANFIS Architecture .....	39
Figure 3.2: Rule selection comparison between CF and ANFIS .....	40
Figure 3.3: Modelling Flowchart of Chapter 3.....	41
Figure 4.1. Break-even Cost for the Top Performing Survivor of the DFDR <sup>+</sup> Procedure (IS 2 Years) .....	72
Figure 5.1: Model Performance Density .....	95
Figure 5.2: Survival Rate for Alternative Means Specifications Across the Markets. .....	98
Figure 5.3: Conditional Variance Survival Proportion Dynamics Across the Markets .....	99
Figure B.1: Average Firing Strength for the $w1/12$ .....	113
Figure C.1. Break-even Cost for the Top Performing Survivor of the DFDR <sup>+</sup> Procedure (IS 1 Year) .....	125
Figure D.1: Model Performance Density for MAE1.....	130
Figure D.2: Model Performance Density for MAE2.....	130
Figure D.3: Model Performance Density for MSE2 .....	131
Figure D.4: Model Performance Density for R2LOG .....	131
Figure D.5: Model Performance Density for QLIKE .....	132
Figure D.6: Conditional Mean Survival Proportion Dynamics Across the Markets for MAE1.....	133
Figure D.7: Conditional Mean Survival Proportion Dynamics Across the Markets for MAE2.....	134
Figure D.8: Conditional Mean Survival Proportion Dynamics Across the Markets for MSE2 .....	134
Figure D.9: Conditional Mean Survival Proportion Dynamics Across the Markets for R2LOG.....	135
Figure D.10: Conditional Mean Survival Proportion Dynamics Across the Markets for QLIKE .....	135

Figure D.11: Conditional Variance Survival Proportion Dynamics Across the Markets for MAE1.....	136
Figure D.12: Conditional Variance Survival Proportion Dynamics Across the Markets for MAE2.....	136
Figure D.13: Conditional Variance Survival Proportion Dynamics Across the Markets for MSE2 .....	137
Figure D.14: Conditional Variance Survival Proportion Dynamics Across the Markets for R2LOG .....	137
Figure D.15: Conditional Variance Survival Proportion Dynamics Across the Markets for QLIKE .....	138

## Acknowledgements

I would like to dedicate a few words to express my gratitude to those who made this writing happen.

My family supported me at every turning point in my life. My dad has been the best advisor throughout my life and my mom is a comprehensive definition of unconditional love. And my beloved beautiful sister has always been the taste of life in our family. My family provided me with the “support points” to map the compassion and I’m wholeheartedly grateful for that.

When I got on board the flight to Glasgow in September 2015 I had no clue about what a PhD is in practice. Professor Georgios Sermpinis has been the one who led me from A to Z on my PhD journey. He has always been there for academic and non-academic supports. Without him being my guarantor, I could not even rent a place to live. Dr Charalampos Stasinakis has been my mentor who always helped me with my work-life balance and not to get too much frustrated. I thank them for their great treat and giving me the opportunity to do this degree.

I find myself struggling often in dealing with human communication conventions. I must thank my supervisors again for their prolonged patience when I had my moments.

I would like to thank Dr Philip Newall who proofread my thesis and helped me a lot with writing texts that are understandable and easy to follow. Moreover, he always had a sympathetic ear to lend when I felt really low.

I thank the Department of Economics at Adam Smith Business School for their support and funding my research expenses. The PhD Workshops held every year have been great settings for presenting and discussing the students’ works. I should also thank Professor Christian Ewald for his comments and ideas to improve the second and third essays in these workshops in 2017 and 2018.

I benefited a lot from expert’s opinion in improving my chapters. My supervisors facilitated this process and introduced the best people in each field to get their opinion. Specifically, I received priceless comments from Dr Thanos

Verousis for the second essay and Dr Ioannis Psaradellis for the third one. I should also thank Dr Jason Laws, Professor Frank McGroarty, Professor Hans-Jörg von Mettenheim, Professor Neil Kellard, and Dr Hans-Georg Zimmermann for their comments on my works in Forecasting Financial Markets conferences in 2016 and 2017.

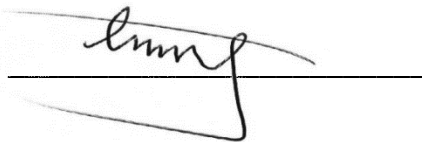
I thank my best friend Arthur Galichère for playing video-games and going to the cinema with me every week over the past two years. I should also thank Moritz Mosenhauer for his kindness and advice on presenting my work. I wish them both great success in their PhD studies.

And finally, last but not the least, I thank Magdalena Hristova and Adalberto Guerra Cabrera for helping me to polish this thesis.

## Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Printed Name: Arman Hassanniakalager

Signature: 



## Abbreviations

<b>ADF</b>	Augmented Dicky-Fuller [test]
<b>AI</b>	Artificial Intelligence
<b>AMH</b>	Adaptive Market Hypothesis
<b>ANFIS</b>	Adaptive Neuro-Fuzzy Inference System
<b>ANN</b>	Artificial Neural Network
<b>AR</b>	AutoRegressive
<b>ARCH</b>	AutoRegressive Conditional Heteroskedasticity
<b>ARV</b>	Adjusted Realized Volatility
<b>BN</b>	Bayesian Network
<b>BNN</b>	Bayesian regularised Neural Network
<b>BRC</b>	Bootstrap Reality Check [test]
<b>CB</b>	Channel Breakout
<b>CF</b>	Conditional Fuzzy
<b>DFDR</b>	Discrete False Discovery Rate
<b>DJIA</b>	Dow Jones Industrial Average
<b>DMA</b>	Dynamic Model Averaging
<b>DMS</b>	Dynamic Model Selection
<b>DRB</b>	Discrete Right-Boundary
<b>EGARCH</b>	Exponential GARCH
<b>EWMA</b>	Exponential Weighted Moving Average
<b>FDP</b>	False Discovery Proportion
<b>FDR</b>	False Discovery Rate
<b>FI-GARCH</b>	Fractionally Integrated GARCH
<b>FIR</b>	Filter Rule
<b>FIS</b>	Fuzzy Inference System
<b>FL</b>	Fuzzy Logic
<b>FR</b>	Fuzzy Rule
<b>FTSE</b>	Financial Times Stock Exchange
<b>FWER</b>	Family-Wise Error Rate
<b>FX</b>	Foreign eXchange
<b>GARCH</b>	Generalized AutoRegressive Conditional Heteroskedasticity
<b>GED</b>	Generalized Error Distribution
<b>HAR</b>	Heterogeneous AutoRegressive
<b>IGARCH</b>	Integrated GARCH
<b>IS</b>	In-Sample
<b>JB</b>	Jarque-Bera [test]
<b>LR</b>	Logistic Regression
<b>MA</b>	Moving Average

<b>MAE</b>	Mean Absolute Error
<b>MAP</b>	Maximum A Posteriori
<b>MHT</b>	Multiple Hypothesis Testing
<b>ML</b>	Machine Learning
<b>MSCI</b>	Morgen Stanley Capital International
<b>MSE</b>	Mean Square Error
<b>NB</b>	Naïve Bayes
<b>OBV</b>	On-Balance Volumes
<b>OFR</b>	Office of Financial Research
<b>OOS</b>	Out-Of-Sample
<b>OP</b>	Ordered Probit
<b>OR</b>	Operational Research
<b>P-P</b>	Philips and Perron [test]
<b>PT</b>	Pesaran-Timmermann [test]
<b>RBF</b>	Radial Basis Function
<b>RM</b>	RiskMetrics
<b>RSI</b>	Relative Strength Indicators
<b>RV</b>	Realized Volatility
<b>RVM</b>	Relevance Vector Machine
<b>RW</b>	Romano and Wolf [test]
<b>S&amp;P</b>	Standard and Poor's
<b>S&amp;R</b>	Support and Resistance
<b>SA</b>	Simple Average
<b>SI</b>	Statistical Inference
<b>SPA</b>	Superior Predictive Ability
<b>SV</b>	Stochastic Volatility
<b>SVM</b>	Support Vector Machine
<b>TGARCH</b>	Threshold GARCH
<b>TVP</b>	Time-Varying Parameter
<b>VaR</b>	Value at Risk

# 1. Introduction

## 1.1 Background and Motivation

Walking down the pavements of financial districts in almost any country on earth could lead you to hear the terms *fundamental* or *technical* analysis. They are the most common approaches to find gainful investment opportunities across capital markets. In fundamental analysis, the expert tries to evaluate the intrinsic value of a security by examining multiple economic or financial factors. Traders using technical analysis inspect historical price charts to predict future market movements, irrespective of fundamental factors. These approaches have historically been conducted by professionals, who spent a notable proportion of their lives gaining experience on the trading floors.

By the digital revolution in the 1970s to 1980s and the advent of advanced learning algorithms in 1990s, humans found themselves competing for jobs with new rivals: *the machines*. Computers can learn and improve their wisdom without the limitation of humans needs or constraints. If in the early days machines were used solely to replace manual labour, today they can beat the best human experts in the games chess and go, and bluff better in poker than humans (Newall, 2013; Newall, 2018). Neuroscience surveys show that the human brain could be thoroughly replicated in terms of computation speed and storage capacity in the digital world by 2020 (see e.g. Markram, 2012). Recent job market studies show that 1 out of 3 UK jobs in the finance industry is at “potential high risk” of automation by 2030 (see among others, Boston Consulting Group, 2015; PwC, 2017). If in the 1980s replacing humans with machines was seen as an entertaining science fiction story like *Blade Runner*, three decades later it is an alert raised by the legendary physicist Stephen Hawking:

"Humans, who are limited by slow biological evolution, couldn't compete and would be superseded [by machines]" (Hawking, 2014).

In financial markets, a professional trader's insights could potentially be replicated by a model running on a computer. Quantitative analysis is the new investment approach where models or algorithms light the path to the highest rewarding investments across multiple asset classes. The models have a mostly

statistical nature and take systematic risk into account. The investment strategies articulated by the quantitative approach can entail a wide range of complexity. Complexity ranges from the automation of simple technical rules to advanced strategies combining quantitative and fundamental analysis: the “*quantamental*” approach.

This thesis brings together several quantitative solutions to financial decision making and analyses their performance through four empirical studies. The thesis adds to the literature of novel approaches to investments, where human expertise is partially or entirely imitated. In all empirical chapters, there are four common elements. First, a time series provides the population for the study. Second, multiple samples are taken from the population as IS and OOS. The number of elements within each sample varies, but the analyses are generally reported on an annual basis. The third common component is the quantitative strategy replicating human wisdom based on the IS observations (versus experience on the market) to predict the unseen cases (future) in the OOS. And the fourth element is a numerical scale to measure and compare model performance over time.

The strategies and applications vary from one chapter to another. The quantitative strategies are chosen to cover a wide range of models used on a regular basis by top-tier investment corporations. The strategies originate from three categories of models: technical trading rules, Machine Learning (ML), and conventional statistical models. These models are deployed for financial predictions under Artificial Intelligence (AI) and Statistical Inference (SI) contexts. In AI, a learning algorithm is used to map an underlying relationship between a set of inputs and outputs by minimizing a loss function. The SI tries to draw a conclusion on the performance of the candidate model through hypothesis testing. In SI, many models are compared to find the genuinely superior ones. The goal is to make sure that findings are robust, rather than being affected by temporary market trends or pure luck.

The design of studies amplifies the difference in performance of alternative models. For the AI surveys, FX and sports betting are chosen as models of high volatility markets, where the potential loss from wrong decisions is high. Such settings can exhibit the superiority of the novel AI models. For the SI surveys,

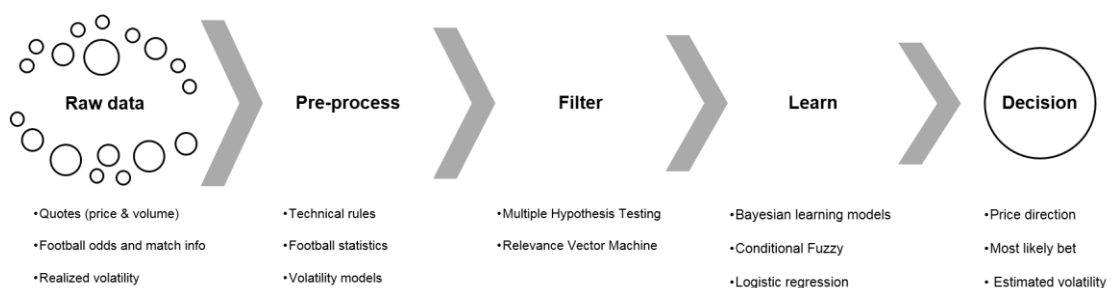
highly significant markets – FX, stocks, and commodities – are studied from around the globe to show the economic significance of market inefficiencies.

## 1.2 Contribution and Structure

The main focus of this research is developing original predictive systems driven by quantitative models. AI is an adaptive solution for making financial predictions by learning from the past to project the future. The AI systems require a training dataset to produce either a set of “*if ... then ...*” rules or optimize a set of parameters to predict a class or point. They are able to provide accurate estimation only when they are given the right amount of information for analysis. An insufficient number of observations causes a poor mapping (under-fitting) for the interaction between the input and output vectors, while an excessive supply of inputs can blur the model’s ability to connect the dynamics of inputs with the outputs (over-fitting). In the era of big data and countless predictive models, dealing with the dimensionality of the dataset is a prerequisite to the success of AI. Reducing the input space to the most informative subset is required for a profitable future of AI systems.

I propose a paradigm where the learning algorithms are coupled with statistical approaches to battle the dimensionality problem. The paradigm is inspired by the concept of deep learning and is altered to detect the patterns in financial markets. Figure 1.1 presents the novel paradigm that can autonomously make investment decisions from raw data. The paradigm accommodates the methodology used in all chapters of this thesis.

**Figure 1.1: Proposed Paradigm for Financial Decision Making with Examples**



**Note:** This system can make decisions by scanning through publicly available raw data. In the Pre-process layer, a large set of potential predictors is generated from the raw data. In the Filter layer, a statistical approach decides on the most informative subset of predictors. This guarantees that a right amount of data is fed to the learning algorithm. In the Learn stage, an AI method is used to detect the patterns between the selected predictors and a target series. The target series can originate from

any topic that a practitioner needs to decide on. The output of the AI model is the decision recommended by the developed expert system.

The rest of this Section discusses the relevance of each chapter to this framework and the novelty of each empirical study. Chapter 2 presents the empirical evidence of superior performance in support of the paradigm in Figure 1.1. The goal of this Chapter is to revisit the validity of technical trading rules and see whether it is possible to generate profit in the FX market using the proposed paradigm. First, a pool of seven thousand eight-hundred forty-six common technical trading rules is generated based on the time series of price and volume quotes over 2010-2016. An SI method – Multiple Hypothesis Testing (MHT) – is used to screen the potential set of outperforming candidates from the technical trading pool. The MHT is a platform that allows performance comparison of alternative candidates while controlling the probability of Type I errors (false positives). Then four Bayesian learning algorithms are used to construct portfolios based on the reduced set of inputs. The findings show the significant difference in trading performance between the proposed framework and the benchmarks in the FX market.

In Chapter 3, the proposed framework is studied with a new set of models in an extremely risky environment. More specifically, a novel class of AI models for the learning component of Figure 1.1 is introduced and applied in predicting football match outcomes. In sports betting, the profit from a bet can exceed 100%, but might also lead to the total loss of the wager should the predicted outcome fail to occur. The raw data consists of the average quotes for winning different types of bets from the bookmakers and games' statistics including goals scored and shots on target. The inputs are processed to form a set of eighty-three explanatory variables based on the odds and recent performance of each playing football team over 2005-2016. In the Filter unit, a Bayesian probabilistic approach (RVM) is used to find the subset of explanatory variables most helpful in predicting the bet outcome. Finally, in the Learn component a novel extension of rule-based models, namely CF inference system, is introduced to extract and apply the rules with the highest level of confidence. This modification to the ordinary Fuzzy Logic (FL) safeguards against under-fitting and over-fitting problems. The results present another successful implementation of the proposed paradigm and the superior performance of the CF algorithm.

Chapters 4 and 5 address Figure 1.1's Filter module. In these chapters, a novel class of MHT is introduced in different applied settings. Chapter 4 presents the concept of DFDR<sup>+/-</sup> as a new MHT method. Then, a pool of twenty-one thousand one hundred ninety-five technical trading rules is analysed for the transaction costs, profitability, and robustness of the rules. The analysis is conducted for twelve stock markets around the globe over the period 2006-2015. The results highlight the role of transaction costs in financial markets, the gainful opportunities available in emerging and frontier markets, and ultimately the role of active trading. The findings are consistent with the theory of Adaptive Market Hypothesis (AMH) introduced by Lo (2004).

Chapter 5 provides a novel application of the DFDR<sup>+/-</sup> method in risk management. In this Chapter, the largest pool of models in the volatility forecasting field<sup>1</sup>, is compiled based on high-frequency time-series over 2013-2017. The models are applied to six assets from stock, currency, and commodity markets. The pool of one thousand five hundred and twelve unique volatility forecasting models is compared to find out whether there is any statistical difference in the predictive ability of volatility models. The models come from twenty classes of four common families of AutoRegressive (AR) models (GARCH, EWMA, SV, and HAR). The results show that advanced specifications provide superior accuracy in almost all markets compared to the three benchmarks.

All in all, this thesis shows how the proposed paradigm in Figure 1.1 can generate superior performance by studying four cases in trading, betting and volatility forecasting. Finally, Chapter 6 delivers several concluding remarks.

---

<sup>1</sup> To my best knowledge as of August 2018.

## 2. Trading the Foreign Exchange Market with Technical Analysis and Bayesian Statistics

### 2.1 Introduction

Technical analysis is the study of past market data in order to forecast the direction of financial asset prices. Its origins can be traced back to the Dow theory in 1900 when Charles H. Dow argued that the financial markets follow repetitive trends. Practitioners apply this principle in practice and many technical trading rules were developed over the next decades aiming to identify the future direction of financial assets. An industry was created based on the application of mathematics in trading. Today thousands of professionals trade financial series with mathematical models.

The most heavily traded assets in financial markets are FX pairs with a turnover of up to \$5.3 trillion daily in 2013 (Jorion, 1996; BIS, 2013). The enormous size of the FX market, the competition among market participants and the advent of technology have led to a continuous search for more advanced and complex trading rules. Researchers and practitioners borrow algorithms from mathematics, physics, genetics and computer science in an attempt to model series that have a non-linear and non-stationary structure. Some apply simple technical rules (Gençay *et al.* 2003; Qi and Wu, 2006; Neely *et al.*, 2009; Cialenco and Protopapadakis, 2011) while others explore complex non-linear models (Neely *et al.*, 1997; Gehrig and Menkhoff, 2006; Gardojevic 2007; Sermpinis *et al.*, 2015). There are also academics that believe FX series follow a random walk and any profitable trading rules are due to luck (Meese and Rogoff, 1983; MacDonald and Taylor, 1994; Kilian and Taylor, 2003).

This Chapter utilizes the latest developments in time-series modelling and statistics in order to discover whether simple technical rules are profitable in FX trading series. It also explores whether it is possible to combine simple technical rules with a set of some of the most up-to-date Bayesian models (RVM, DMA, DMS, and BNN) and derive superior trades.

For this purpose, seven thousand eight hundred forty-six technical rules are generated and applied to three exchange rates (EUR/USD, GBP/USD, and



USD/JPY). Next, the genuinely profitable trading rules are identified based on the Romano *et al.* (2008) test combined with the balancing procedure of Romano and Wolf (2010). These profitable rules are then combined with NB, RVM, DMA, DMS and BNN. It is worth noting that the RVM, DMA, DMS and BNN have not been used in a trading application<sup>2</sup>. The results show that superior trading performance is achievable by combining a data snooping procedure and Bayesian learning models. This Chapter finds that BNN, DMA, and DMS have the highest performance across the study periods.

The motivation for this study derives from four sources: the AMH, the contradicting reports on the value of technical analysis in trading, the popularity of Bayesian techniques in financial forecasting, and the increased use of computational techniques in trading. The AMH has three main principles: traders need to be adaptive, the performance of trading models varies through time and in competitive environments the opportunities for profits are scarce. In other words, in highly efficient markets simple trading strategies have small power and traders need to seek complex statistical methods that are adaptive to the changing environment. The FX market – the biggest capital market – is most competitive and it is heavily affected by the intervention of central banks. It is interesting to check the effectiveness of simple trading rules in this environment and if possible to generate Bayesian combinations of simple rules that can beat a market. I also examine if the performance of the trading models varies through time and whether their profitability is less in “popular” exchange rates, as the AMH proposes.

Technical analysis is considered a universal trading practice across different markets (Blume *et al.*, 1994). Although theories around technical analysis vary, all of them are based on the idea of the recurrent nature of patterns in the securities’ price charts. Chartists believe that understanding these patterns can facilitate the prediction of future prices (Fama, 1965). This approach to prediction

---

<sup>2</sup> To the best of my knowledge, RVM has only one related application (Fletcher *et al.*, 2009) on FX carry trade. In this Chapter the RVM is used as a part of a set of AI models and its individual performance is not assessed. The BNN also has only one application in financial forecasting in Ticknor (2013). In his study, BNN is not evaluated in trading terms. I did not identify any related trading application of DMA and DMS although there are several studies with them in financial and economic modelling (such as Koop and Korobilis, 2012; and Byrne *et al.*, 2016).

of financial markets can be traced back to Dow theory. The theory argues that the average values represent net interactions of all market participants over day-to-day activities and discount all kind of news and events, even the unpredictable ones. It proposes three bands of trends known as primary, secondary and minor trends. The primary trends are major market movements known as bull and bear market. The secondary trend represents the corrections and recoveries over bull and bear markets respectively. Finally, the minor trends are daily meaningless fluctuations (Edwards *et al.*, 2007). Several studies, such as Sweeney (1988), Brock *et al.* (1992) and Blume *et al.* (1994) demonstrate the utility and the profitability of technical analysis in financial markets. In these studies, a large universe of simple trading rules is generated, and their average performance is evaluated on stocks or stocks indices over a large period of time. Gençay (1998) uses technical rules as inputs to Artificial Neural Networks (ANNs) and generates profitable models, while Allen and Karjalainen (1999) use a genetic algorithm to identify profitable technical trading rules for the Standard and Poor's (S&P) 500 index. Although these preliminary studies seem promising, they ignore the data snooping bias.

Data snooping occurs when a given dataset is used more than once for purposes of inference and model selection (White, 2000). This bias is prominent in trading applications where researchers rely on the same data set to test the significance of different trading rules individually. These individual statistics are generated from the same dataset and relate to each other. White (2000) formalises this bias and introduces the Bootstrap Reality Check (BRC), which considers the dependence of individual statistics. The introduction of BRC test allowed researchers to revisit technical analysis from a new angle. Sullivan *et al.* (1999) claim that, based on the BRC test, technical analysis has no value on Dow Jones Industrial Average (DJIA) index. Hansen (2005) argues that the BRC is too conservative and checks only whether there is any significant model. The BRC does not identify all such models. As a solution, Hansen (2005) introduces the Superior Predictive Ability (SPA) test, which is less conservative and seems more powerful (Hansen and Lunde, 2005). Hsu and Kuan (2005) study technical rules after taking into account data snooping with the SPA test and claim that it is possible to beat the market with complex rules. Romano and Wolf (2005) and Hsu *et al.* (2010) improve the BRC and the SPA test respectively and introduce stepwise procedures

Step-BRC and Step-SPA. These tests can identify all possible significant models. Further improvements in MHT procedures are made by Romano and Wolf (2007), Romano *et al.* (2008), Bajgrowicz and Scaillet (2012) and Hsu *et al.* (2014). The trend in recent data snooping literature is to relax the statistics by controlling the probability of making multiple false rejections (falsely “found” profitable strategies) and at the same time improve the efficiency of the tests. This is beneficial in trading applications, where large groups of technical rules are under study and the ability to make true rejections is the main concern. Based on the latest tests, Romano and Wolf (2007), Romano *et al.* (2008), Bajgrowicz and Scaillet (2012) and Hsu *et al.* (2014) conclude that it is possible to identify genuinely profitable trading rules by using an efficient MHT procedure. However, the same studies argue that the profit margins are small, and the trading performance varies through time.

In FX market specifically, Gehrig and Menkhoff (2006) argue that technical analysis has by far the greatest importance for FX trading. Gençay *et al.* (2003) generate positive annualized returns on four currency pairs with a real-time trading based on simple exponential Moving Average (MA) models. However, Cialenco and Protopapadakis (2011) argue that simple trading rules do not report statistically significant profitability in fourteen currencies. Meese and Rogoff (1983), Baillie and Bollerslev (1989), and Chinn and Meese (1995) claim that major exchange rates follow a random walk (at least in the short-run). Taylor (1992) reports that 90% of the chief FX dealers based in London place some weight to technical analysis in their decision processes. Yilmaz (2003) suggests that FX prices do not always follow a martingale<sup>3</sup> process, especially during the periods of central banks interventions. Yang *et al.* (2008) argue that martingale behaviour cannot be rejected for major exchange rates. Contrary to these, MacDonald and Taylor (1994) develop a monetary model which outperforms the random walk for the GBP/USD exchange rate over short-run periods. Kilian and Taylor (2003) find strong evidence of predictability over horizons of 2 to 3 years with a similar model, but not over shorter horizons. Hsu *et al.* (2010) and Hsu *et al.* (2014) argue that

---

<sup>3</sup> Martingale corresponds a sequence of random variables where the expected value for the next observation is equal to the present one or  $E(\zeta_{t+1}|\zeta_1, \dots, \zeta_t) = \zeta_t$ .

technical analysis can beat the FX market. The same statement is made by Neely and Weller (2013) who add that traders need to be adaptive in their portfolios.

Developments in statistics and computer science offer new potentials for wealth management. The developments include the advent of new tools in the fields of MHT and AI. Chui *et al.* (2016) and the Boston Consulting Group (2015) estimate that by 2025 the field of wealth management will be dominated by ML. In academia, there is a plethora of studies in this field. Gençay (1998), Fernández-Rodríguez *et al.* (2000), Jasic and Wood (2004), Gradojevic (2007), Sermpinis *et al.* (2013), and Sermpinis *et al.* (2015) apply ANNs – a form of non-linear regression algorithms – to the task of forecasting and trading financial series with some success. Alvarez-Diaz and Alvarez (2003), Pai *et al.* (2006), and Huang *et al.* (2010) develop models inspired by the evolution of species to financial forecasting with good results. Allen and Karjalainen (1999) use a genetic algorithm to identify profitable technical trading rules for the S&P 500 index. Lin and Pai (2010), Bekiros (2010) and Gradojevic and Gençay, (2013) apply FL in order to generate trading signals. Other studies, such as Ticknor (2013) and Gramacy *et al.* (2014), use Bayesian Statistics in financial forecasting problems. The literature in the area is extensive and promising. In the papers that have a trading application (see among others, Jasic and Wood, 2004; Gradojevic and Gençay, 2013) the proposed complex models significantly outperform simple trading rules. An explanation can be offered by the AMH which argues that complex models can survive better in informative markets.

In a nutshell, the literature in technical analysis, data snooping and computational applications in trading, is wealthy and contradicting. Studies that do not consider the data snooping bias and involve models that require parametrization should be treated with scepticism. The data snooping bias should be examined with recent related tests that are not strict. Computational techniques seem able to generate profitable trades. However, it is not clear from the previous studies if computational models can outperform technical analysis, as the AMH claims.

## 2.2 Methodology

In this study, a large set of technical trading rules on FX data is generated. The genuine profitable rules are identified with the Romano *et al.* (2008) test as modified based on the balancing procedure of Romano and Wolf (2010). Then, the profitable rules are combined with an NB, an RVM, a DMA, a DMS, and a BNN. The Bayesian methods are chosen from a wide range of complexity. Such choice allows quantifying the differences in performance of the learning models in practice. The next sections contain a short description of the data snooping procedure and the Bayesian techniques<sup>4</sup>.

### 2.2.1 Data Snooping Test

At the first stage for the modelling, the genuinely profitable trading rules are identified from a pool of 7846 technical rules. For this purpose, the Romano *et al.* (2008) test is combined with the balancing procedure of Romano and Wolf (2010). The benefits of the proposed approach are threefold. Firstly, it considers different measures of errors. Secondly, it is balanced since each individual hypothesis is treated fairly. Finally, it involves a resampling and subsampling approach that considers the dependence structure of the individual test statistics. These facts make it highly applicable in trading applications and more efficient compared to the Step-BRC and Step-SPA tests (Romano and Wolf, 2005; Hsu *et al.*, 2010).

The data snooping test is an MHT procedure in which a set of models are tested to identify the statistically different ones. As in any statistical test, there is the chance that a hypothesis is falsely rejected (Type I error). Familywise Error Rate (FWER) is the probability of having at least one false rejection. Traditional data snooping tests are too strict as they are attempting to control (asymptotically) the FWER. If the number of hypotheses is very large (as in this Chapter's case), it is very difficult to make true rejections. In the asset

---

<sup>4</sup> These algorithms are characterized by their complexity (except for NB). For the sake of space and as their mathematical derivation already exist in the relevant literature, I present the general framework. For the data snooping procedure, the reader is referred to Romano *et al.* (2008) for the FWER control with one sided setup and for the balancing procedure to Romano and Wolf (2010). A detailed description of RVM is provided by Tipping (2001) while a complete mathematical derivation of DMA and DMS is provided by Raftery *et al.* (2010). The Bayesian training procedure of BNN is described in detail in Ticknor (2013).

management industry, professionals diversify their risk by investing in a large portfolio of models. The performance of any bad model is diluted by the much larger set of profitable rules.  $k$ -FWER determines the probability of having at least  $k$  false rejections. The data snooping approach of Romano *et al.* (2008) tries to control the  $k$ -FWER<sup>5</sup>.

Let me consider a set of  $\mathcal{S}$  trading strategies over  $T$  sample periods. For each trading strategy  $s$  (where  $s = 1$  to 7846), the aim is to test the hypothesis that a model  $s$  beats a benchmark ( $\varsigma$ ) in terms of profitability. Let me define  $\mu_s$  as the unconditional average profit of the strategy  $s$  and  $\theta_s = \mu_s - \mu_\varsigma$  as its difference from the benchmark. The null hypothesis is  $H_{0,s}: \theta_s \leq \theta_\varsigma$ , while the alternative is  $H_{1,s}: \theta_s > \theta_\varsigma$ . This setting tests the hypothesis that the technical rules have an equal or worse profitability compared to the benchmark  $\varsigma$ . The test statistic is set as:

$$Z_{T,s} = \frac{\bar{r}_{T,s} - \bar{r}_{T,\varsigma}}{\hat{\sigma}_{T,s}} \quad (2.1)$$

where the historical mean  $\bar{r}_{T,s}$  and standard deviation  $\hat{\sigma}_{T,s}$  in the (2.1) are given by:

$$\bar{r}_{T,s} = \frac{1}{T} \sum_{t=1}^T \hat{r}_{t,s}, \quad (2.2)$$

$$\hat{\sigma}_{T,s} = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (r_{t,s} - r_{t,\varsigma})^2}. \quad (2.3)$$

The  $k$ -FWER is controlled through the one-sided setup of the  $k$ -StepM method of Romano and Wolf (2005). Firstly, the strategies are sorted in a descending order based on the test statistics. After this is done, if  $b_k$  is the  $k$ -largest test statistic, then  $Z_{T,b_1} \geq \dots \geq Z_{T,b_S}$ . Next, the  $k$ -th largest test statistic and the  $1 - \alpha$  (where  $\alpha$  is the significance level) percentile of its sampling distribution are estimated. The individual hypotheses outside the confidence region are rejected. For the hypotheses not rejected, the process is repeated until the number of rejections is

---

<sup>5</sup> Chapter 4 provides a comprehensive discussion on the MHT and introduces a novel approach to estimate and control the Type I error.

smaller than the desired  $k$ . For more details on the Step-M methods and the relevant bootstrap approach, see Romano *et al.* (2008) or Mazzocco and Saini (2012). In order to control the  $k$ -FWER, the innovations of Romano and Wolf (2010) are followed. They introduce an asymptotically balanced method that controls the average number of false rejections. Implicitly this approach considers the dependence structure of the individual test statistics, which leads to a more efficient control of false null hypotheses (Type II error<sup>6</sup>). In this application, most technical trading rules have some form of weak dependency. (For instance, two MA cross-over strategies with different fast-MA of 2 and 5 periods but a similar slow-MA of 75 periods).

The selection of  $k$  depends on the problem under study and the practitioner's approach. If  $k$  is 1, the method can be overly conservative and inefficient<sup>7</sup>. For this study, the  $k$  is set to 39 (roughly 0.5% of the 7846 technical rules under study<sup>8</sup>). As a benchmark to the data snooping test ( $\varsigma$ ), a basic random walk model is applied since major exchange rates are widely suggested to follow a random walk (see among others, Meese and Rogoff, 1983; Baillie and Bollerslev, 1989; and Chinn and Meese, 1995).

### 2.2.2 RVM

The RVM approach proposed by Tipping (2001), seeks to find the most effective inputs based on probabilistic approaches to classification and regression problems. Throughout this process, the determined effective points are defined as relevance vectors. This Section summarizes the RVM structure.

---

<sup>6</sup> Type II error corresponds to cases where a true alternative hypothesis is falsely not rejected. Since the goal of the MHT is controlling Type I error, the risk of ignoring the significant rules increases when the data snooping procedure is conservative. Table C.3 shows how excess conservativeness leads to Type II error where true discoveries are reported insignificant.

<sup>7</sup> Appendix C.1 provides evidence based on Monte Carlo simulations how  $k = 1$  can lead to very poor detection of significant models.

<sup>8</sup> The choice of 0.5% is based on *approximating* the set of initial rejections (including both true and false discoveries) with the top 5 percent of trading rules and allowing 10% of rejections to be the Type I error. This *approximation* can be improved by alternative statistical approaches to the rejections based on the test statistics and the bootstrap  $p$ -value (presented in Chapter 4). However, this *approximation* is chosen to find the most profitable trading rules.

Assuming a supervised learning framework, I define a dataset  $D$  with  $v$  predictors and  $T$  training points, an input series set  $\mathbf{x} = \{x_i: i = 1, \dots, T\}$  and a target series set  $\mathbf{y} = \{y_i: i = 1, \dots, T\}$ . The general predictive formulation can be specified as:

$$y_i = f(x_i) + \varepsilon_i \quad (2.4)$$

where  $\varepsilon_i$  is the zero-mean Gaussian error term with distribution  $\varepsilon_i = y_i - \hat{y}_i \sim \mathcal{N}(0, \sigma^2)$ ,  $\hat{y}_i$  is the target point forecast, and  $f$  is the transfer function.

Given the basis function set  $\boldsymbol{\varphi}(\mathbf{x})$  and the weight vector  $\mathbf{w}$ , the RVM's prediction under the linear model assumption can be expressed as:

$$\hat{\mathbf{y}} = f(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^v w_j \varphi_j(\mathbf{x}) + w_0 \quad (2.5)$$

where  $\boldsymbol{\varphi}(\mathbf{x}) = [1, K(\mathbf{x}, x_1), \dots, K(\mathbf{x}, x_T)]'$ ,  $w_0$  is the bias, and  $\mathbf{w} = [w_1, \dots, w_v]$ .

In the context of RVM, Radial Basis Function (RBF) is mostly considered as the basis function  $K$ . This is due to its simplicity and superior optimization performance (Park and Sandberg, 1991). Subsequently, the multivariate Gaussian likelihood of the dataset can be written as:

$$Pr(\mathbf{y}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{\|\mathbf{y} - \Phi\mathbf{w}\|^2}{2\sigma^2}\right) \quad (2.6)$$

where  $\Phi$  is the  $T \times (T + 1)$  'design' matrix with  $\Phi_{nm} = K(x_n, x_{m-1})$  and  $\Phi_{n1} = 1$ .

Over-fitting can be expected in the maximum-likelihood estimation of  $\mathbf{w}$  and  $\sigma^2$  in Eq. (2.6). To overcome this, Tipping (2001) recommends setting prior constraints on parameters  $\mathbf{w}$  by adding a complexity term inspired by the traditional margin concept of Support Vector Machine (SVM) modelling. Gaussian priori in RVM context for an individual  $w_j$  can be expressed as:

$$Pr(w_j|\alpha_j) = \left(\frac{\alpha_j}{2\pi}\right)^{1/2} \exp\left(-\frac{\alpha_j w_j^2}{2}\right) \quad (2.7)$$



Similarly, for the whole set of  $\mathbf{w}$ :  $Pr(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^T \mathcal{N}(w_i|0, \alpha_i^{-1})$ , where  $\boldsymbol{\alpha} = [\alpha_0, \dots, \alpha_T]'$  is a hyperparameter vector governing the prior defined over the weight  $\mathbf{w}$  to control deviation of each  $w_j$  from the zero mean.

Given priori information controlling the generalisation ability and the likelihood distributions, applying Bayes' rule generates the posterior over  $\mathbf{w}$  as:

$$Pr(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = \frac{Pr(\mathbf{y}|\mathbf{w}, \sigma^2)Pr(\mathbf{w}|\boldsymbol{\alpha})}{Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)} \quad (2.8)$$

In the case of a multivariate Gaussian distribution, the posterior takes the following form:

$$Pr(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (2.9)$$

The covariance and the mean of the distribution are estimated respectively by the following analytical solution of Eq.s (2.10 and 2.11):

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}'\mathbf{B}\boldsymbol{\Phi} + \mathbf{A})^{-1} \quad (2.10)$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}'\mathbf{B}\mathbf{t} \quad (2.11)$$

where  $\mathbf{A} = (\alpha_0, \dots, \alpha_T)$  and  $\mathbf{B} = \sigma^{-2}\mathbf{I}_T$ .

To estimate the weights, the missing set  $\boldsymbol{\alpha}$  in the above equations is treated as a hyperparameter. Therefore, the relevance vector learning model approximates the mode for the hyperparameter posterior i.e. maximization of  $Pr(\boldsymbol{\alpha}, \sigma^2) \propto Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)Pr(\boldsymbol{\alpha})Pr(\sigma^2)$  given  $\boldsymbol{\alpha}$  and  $\sigma^2$ . Assuming uniform hyperparameters, the model optimization can be thought equivalent to the maximization of  $Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ . By Integrating out the weights, the following is derived:

$$Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \int Pr(\mathbf{y}|\mathbf{w}, \sigma^2) Pr(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \quad (2.12)$$

where  $Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$  can be computed by the following equation:

$$Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-T/2} |\mathbf{B}^{-1} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}'|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'(\mathbf{B}^{-1} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}')^{-1}\mathbf{y}\right\} \quad (2.13)$$

The marginal likelihood for hyperparameters in the Gaussian distribution form is given by:

$$Pr(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \mathcal{N}(0, \mathbf{B}^{-1} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}') \quad (2.14)$$

The estimation of the above hyperparameters is conducted through an iterative procedure similar to the gradient ascent on the objective function for Maximum A Posteriori (MAP) estimate of the weights (for more details refer to Ghosh and Mujumdar, 2008; and Candela and Hansen, 2004). The numerical approximation is adopted because there is no closed form solution. The MAP estimation is dependent on the hyperparameters  $\boldsymbol{\alpha}$  and  $\sigma^2$  in other words  $\mathbf{A}$  and  $\mathbf{B}$  in Eq.s (2.10 and 2.11).

Following Tipping (2001) the solution to Eq.s (2.10 and 2.11) is estimated through differentiating and setting Eq. (2.14) to zero. After rearranging we have:

$$\alpha_m^{new} = \frac{\gamma_m}{\mu_m^2} \quad (2.15)$$

where  $\mu_m$  is the  $m$ -th posterior mean-weight from the equation set and  $\gamma_m \equiv 1 - \alpha_m \Sigma_{mm}$ .

The  $\Sigma_{mm}$  is the  $m$ -th diagonal element of the covariance  $\Sigma$  matrix calculated by the updated  $\boldsymbol{\alpha}$  and  $\sigma^2$ . Parameter  $\gamma_m$  is interpreted as the degree to which associated  $w_m$  is well-determined by the training data (MacKay, 1992). When the fit is not appropriate, the  $w_m$  is constrained by priori with small  $\sigma_m^2$ . For example, for a high value of  $\alpha_m$ ,  $\Sigma_{mm}$  will tend to  $\alpha_m^{-1}$  and consequently  $\gamma_m$  approaches zero. On the other hand, when the fit is good,  $\alpha_m \approx 0$ , this leads to  $\Sigma_{mm} \approx 0$ , and finally  $\gamma_m \approx 1$ . Consequently, the range for  $\gamma_m$  is  $[0,1]$ . For the other hyperparameter  $\sigma^2$  differentiation results in the update of the noise variance estimation as:

$$(\sigma^2)^{new} = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{T - \Sigma_m \gamma_m} \quad (2.16)$$

The learning process advances by reestimating the hyperparameters and updating the mean and covariance of the posterior in each iteration. This continues until the convergence is met at an iteration step or until the

incorporated stop criteria are activated to avoid reaching redundant loops. In practice during the iterative update of the hyperparameters, many  $\alpha_j$ s approach infinity. In that way  $w_j$ s tend to form a delta function around zero. Consequently, many elements in  $\mathbf{w}$  and associated elements in  $\boldsymbol{\varphi}(x)$  would be discarded from the operational model. The remaining basis functions that are associated with training points within the sample dataset produce a sparse solution for the RVM model. These remaining examples are the so-called relevance vectors. Tipping (2001) claims that the above predictive estimations are found to be robust by most empirical evidence. The predictive distribution for a given new point  $x_*$  complemented by a  $y_*$  class label is given by:

$$Pr(y_*|x_*, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) = \int Pr(y_*|x_*, \mathbf{w}, \sigma_{MP}^2) Pr(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) d\mathbf{w} \quad (2.17)$$

The Gaussian form is expressed as:

$$Pr(y_*|x_*, \boldsymbol{\alpha}_{MP}, \sigma_{MP}^2) = \mathcal{N}(\hat{y}_*, \sigma_*^2) \quad (2.18)$$

where  $\hat{y}_* = \boldsymbol{\mu}'\boldsymbol{\Phi}(x_*)$  is the mean estimate of the target and  $\sigma_*^2 = \sigma_{MP}^2 + \boldsymbol{\Phi}(x_*)'\boldsymbol{\Sigma}\boldsymbol{\Phi}(x_*)$  is the corresponding uncertainty. The  $\boldsymbol{\alpha}_{MP}$  and  $\sigma_{MP}^2$  are the most probable hyperparameter values obtained from Eq. (2.13).

The predictive mean is generated through the reduced basis function and the input explanatory variables. The predictive variance confirms that the OOS prediction is consistently higher than the IS one due to extra uncertainty caused in the process of the weights prediction.

### 2.2.3 DMA and DMS

Financial trading series are dominated by structural breaks. Models with fixed coefficients work only for short periods. Time-Varying Parameter (TVP) models consider the parameters as a function of time and are estimated using state-space methods such as Kalman filter. Despite the benefits of the TVP models over static methods, the assumption is that the initial set of explanatory variables remains relevant over time. This can be undesirable in real environment applications.

The DMA proposed by Raftery *et al.* (2010) allows selecting different subsets of explanatory variables over time along with variable coefficients. Consider a candidate input set  $u = 1, \dots, U$ , then the state-space model at time  $t = 1, \dots, T$  for the dependent variable  $y_t$  can be presented under observational and state equations as:

$$y_t = F_t^{(u)'} \zeta_t^{(u)} + \varepsilon_t^{(u)}, \quad (2.19)$$

$$\zeta_t^{(u)} = \zeta_{t-1}^{(u)} + \eta_t^{(u)}, \quad (2.20)$$

$$\begin{pmatrix} \varepsilon_t^{(u)} \\ \eta_t^{(u)} \end{pmatrix} \sim \mathcal{N} \begin{pmatrix} R_t^{(u)} & 0 \\ 0 & V_t^{(u)} \end{pmatrix}, \quad (2.21)$$

where  $F_t^{(u)}$  in Eq. (2.19) is a subset from the  $\nu$  potential predictors at each time. The  $\zeta_t^{(u)}$  is a  $p \times 1, p \leq \nu$  vector of time-varying regression coefficients evolving over time by Eq. (2.20). From the specification provided, it is immediately visible that the total number of candidate models is  $U = 2^\nu$ . Unless  $\nu$  is very small, updating the parameters becomes demanding and computationally very slow using a full Bayesian approach. Raftery *et al.* (2010) approximates the solutions of Eqs (2.19 to 2.21) and thus makes the algorithm more efficient. However, the computational burden still increases exponentially when  $\nu$  is large. This makes DMA impractical with standard computer processing when  $\nu$  is larger than 20.

The DMA averages the forecasts across candidate combination of models based on predictive likelihood through a recursive updating scheme. The predictive likelihood estimates the ability of model  $u$  to predict  $y_t$ . The models containing better predictors receive higher predictive likelihood and are associated with higher weights in the averaging process. Respectively, at each time  $t$  two vectors of weights for the model  $u$  are calculated as  $\omega_{t|t-1,u}$  and  $\omega_{t|t,u}$ . The first quantity denotes the weight of a specific model given information available at time  $t - 1$ , while the latter one represents the dedicated weight to the specific model after the model update at time  $t$ . The DMS makes the prediction based on the highest value of weight which is calculated through the updating process. This can be mathematical expressed as:

$$\omega_{t|t,u} = \frac{\omega_{t|t-1,u} L_u(y_t|y_{1:t-1})}{\sum_{l=1}^U \omega_{t|t-1,l} L_l(y_t|y_{1:t-1})}, \quad (2.22)$$

where  $L_u(y_t|y_{1:t-1})$  is the predictive likelihood measured by the realized value of  $y_t$ . By using a forgetting factor  $\delta$ , as suggested by Raftery *et al.* (2010), the weights for the following period are formulated as:

$$\omega_{t+1|t,u} = \frac{\omega_{t|t,u}^\delta}{\sum_{l=1}^U \omega_{t|t,l}^\delta}. \quad (2.23)$$

The  $\delta$  controls the ‘forgetting’ of the entire model set and it can take values in the range of  $0 < \delta \leq 1$ . Raftery *et al.* (2010) suggest  $\delta = 0.99$  as a benchmark, while Koop and Korobilis (2012) recommend  $\delta \in [0.95, 0.99]$ . The recursive calculation starts with a non-informative choice for the initial weight  $\omega_{0|0,u} = \frac{1}{U}$  for  $u = 1, \dots, U$ . The other approximation is used in the estimation of the  $V_t^{(u)}$ . The second forgetting factor,  $\lambda$ , explains the information loss over time. Representing the variance estimator  $\zeta_t^{(u)}$  by  $C_t^{(u)}$ , the conditional variance,  $V_t^{(u)}$  (there is no need to be estimated for each individual model), is calculated as:

$$V_t^{(u)} = (1 - \lambda^{-1}) C_{t-1}^{(u)}. \quad (2.24)$$

In other words, the  $\lambda$  controls the amount of shock affecting the coefficients  $\zeta_t^{(u)}$ . Identical to  $\delta$ ,  $\lambda$  may also take values near to one. This determines the rate of which information loses effect on the model coefficients. Here, it should be noted that by setting  $\delta = 1$ , the DMA is transformed to a TVP model with no change in the subset selection over time. Additionally, by setting  $\delta = \lambda = 1$ , the DMA is simplified to conventional Bayesian Model Averaging with no time-varying characteristic.

The term “forgetting factors” stems from the fact that observations at  $j$  periods ago have a contribution with factor  $\lambda^j$  to the model. As a simple analogy, in the case of having  $\lambda = 0.99$ , it takes 69 periods for the shock from each observation to lose half of its effect on the coefficients. The half-life of the shock to the model is reduced to 14 periods for  $\lambda = 0.95$  and further to 6 in the case of  $\lambda = 0.90$ . The values of the forgetting factors can considerably affect the way models react to the changes of the environment. Various surveys recommend the

direct use of  $\delta = \lambda = 0.99$  as a benchmark (Raftery *et al.*, 2010; Aye *et al.*, 2015). Koop and Korobilis (2012) argue that the performance of competing models with different forgetting factors are robust and perform efficiently. They also conduct a sensitivity analysis for the parameters that shows that the best OOS forecasting results are obtained by setting  $\delta = 0.95$  and  $\lambda = 0.99$ . The study of Koop and Korobilis (2012) is conducted on macroeconomic data and indicates that appropriate selection of parameters under volatile conditions can enhance the predictive ability of the DMA and DMS models. In this study, a wider variety of values e.g.  $\{0.90, 0.95, 0.99, 1\}$  for the parameters are experimented to accommodate a more rapid update of the model specification. This choice accommodates the dynamics and nonlinearities of the market. The wide range for parameters also replicates the behaviour of expert traders on the market floor that constantly revise their trading strategy and if necessary rapidly switch from one approach to another<sup>9</sup>.

#### 2.2.4 BNN

BNN is a specific extension of ANNs that are a class of non-linear models inspired by the work and functioning of biological neurones. In the most common set-up, an ANN has at least three layers. The first layer is called the input layer (where the technical rules are fed). The last layer is called the output layer (where the forecasted value is extracted). An intermediary layer of nodes, the hidden layer, separates the input from the output layer. The number of nodes in the hidden layer controls the complexity the model is able to fit. In addition, the input and hidden layer contain an extra node called the bias node. This node has a fixed value of one and has the same function as the intercept in traditional regression models. Normally, each node of one layer has connections to all the other nodes of the next layer.

The training of the network is to adjust its weights so that the network maps the input value of the training data to the corresponding target value. It begins with randomly chosen weights and proceeds by applying a learning algorithm. The most common procedure is the backpropagation of errors (Shapiro, 2000) which

---

<sup>9</sup> The reported results in Section 2.3 are based on the best performance measured in the IS. The IS in the DMA and DMS cases is the first 80% of observations.

looks for the minimum of the error function (commonly the Mean Squared Error (MSE) between the actual and forecasted values) in weight space using the method of gradient descent.

Ticknor (2013) modifies the training procedure by applying Bayesian regularisation which trains the ANN based on:

$$\Omega = \gamma_1 E_{SE} + \gamma_2 E_{we} \quad (2.25)$$

where  $E_{SE}$  is the sum of the squared errors,  $E_{we}$  is the sum of the squared network weights and  $\gamma_1$  and  $\gamma_2$  are objective function parameters. In this framework, the ANN's weights are considered random variables and their density function based on the rule of Bayes is:

$$Pr(w|D, \gamma_1, \gamma_2, M) = \frac{Pr(D|w, \gamma_1, M)Pr(w|\gamma_2, M)}{Pr(D|\gamma_1, \gamma_2, M)} \quad (2.26)$$

where  $w$  is a vector of the network weights,  $D$  is a vector with the dataset (technical rules in this Chapter's case) and  $M$  is the underlying model (the ANN in this case). Based on Forsee and Hagan (1997), the optimization of parameters  $\gamma_1$  and  $\gamma_2$  requires solving a Hessian matrix based on the Levenberg-Marquardt training algorithm. In order to protect the ANN from over-fitting, the early stopping procedure in the IS is applied.

In BNN, overly complex models are penalized as unnecessary linkage weights and are effectively driven to zero. Burden and Winkler (2009) argue that the network calculates and trains on the nontrivial weights which converges to a constant as the network grows. Parsimonious ANNs limit the training time and the danger of over-fitting. Additionally, they do not require the validation step which is otherwise necessary on the traditional back-propagated ANNs.

### 2.2.5 NB

The RVM models the posterior  $Pr(y|x)$  from the attribute variable set  $x$  to the class label set  $y$ . In the context of probabilistic classification, this approach is termed as discriminative learning. In discriminative classifiers, all training observations from any class  $y_i$  are considered in establishing the model. Despite

evidence in favour of discriminative classification (Vapnik, 1998), there is a reverse approach to the probabilistic classification regarded as generative learning. The generative classifiers learn the joint probability  $Pr(x, y)$ , by using Bayes rules to calculate  $Pr(y|x)$ . Then, a classification model is obtained, which classifies each data point to the label  $y$  with the highest posterior probability. Generative learning is particularly useful, when there is missing information in the dataset.

The NB is a simple classifier that allocates each point of the dataset to the most likely class according to the generative approach. The model is named naïve because of its simplifying assumption that all variables  $x_i$  are conditionally independent for a certain class  $y_0$ . For a test sample with attribute variables  $x = x_0$  of  $v$  dimension and class label  $y = y_0$ , the probability of each class can be calculated by the observed values of the predictive attributes  $x_{j,t}$ ,  $j = 1, \dots, v$ ;  $t = 1, \dots, T$ . By using the Bayes rule, the posterior can be calculated as:

$$Pr(y = y_0 | x = x_0) = \frac{Pr(y=y_0)Pr(x = x_0 | y = y_0)}{Pr(x=x_0)} \quad (2.27)$$

The predicted label is the most probable class given by (2.27). Under the class-conditional independence assumption, I have:

$$Pr(x = x_0 | y = y_0) = \prod_{i=1}^v Pr(x_i = x_{i,0} | y = y_0) \quad (2.28)$$

The conditional distribution  $Pr(x = x_0 | y = y_0)$  may take a multinomial (Gaussian) form for discrete (continuous) variables. Based on the training dataset and plugging the empirical probabilities in Eq.s (2.27 and 2.28), it is easy to make a natural classification as the naïve benchmark.

In this Chapter, NB is used as a benchmark for other types of Bayesian probabilistic models. The attribute variables are the signals generated by the trading rules from a set of  $x \in \{-1, 0, 1\}$ . This set represents short, hold or long positions respectively. Similarly, the class label is the one-step-ahead direction of the market change. For example, a class label  $y \in \{-1, 0, 1\}$  represents the fall, no change or rise respectively of the market in the next period.



## 2.3 Empirical Section

### 2.3.1 Dataset

The proposed methodology is applied to the daily price (open, high, low, and close) and volume series for EUR/USD, GBP/USD, and USD/JPY exchange rates. The period under study is the start of 2010 until the end of 2016 and it is divided into four trading exercises. In each trading exercise, the first three years act as IS and the following year as OOS (i.e. in the first exercise the IS covers the years 2010 until 2012 and the year 2013 acts as OOS).

At the first stage, 7846 simple trading rules are generated for each of the three exchange rates at the IS periods of the four exercises. The trading rules consist of Filter Rules (FIRs), MAs, Support and Resistance levels (S&Rs), Channel Breakouts (CBs), and On-Balance Volume indicators (OBVs). It is the same set of rules applied in the studies of Sullivan *et al.* (1999) and Bajgrowicz and Scallet (2012). For a description of these rules see Appendix A.1. All trading rules are generated through the logarithmic returns of the exchange rates. The summary statistics of the logarithmic returns on daily close for the exchange rates under study are presented in the Table 2.1.

[Table 2.1]

All series exhibit positive kurtosis while the skewness is mixed but generally close to zero. The Jarque and Bera (JB) (1980) test reveals that the return series do not follow a normal distribution, while the Augmented Dicky-Fuller (ADF) (1979) shows that they are stationary. Each trading rule generates a daily trading signal for the relevant exchange rate and IS period. The trading signal can be long (buy), short (sell) or hold (no action). Based on these signals, the trading performance of each of the 7846 rules is generated. As transaction costs, I consider three basis points per trade based on industry rates<sup>10</sup> and academic literature (see among others, Neely and Weller, 2003; Gradojevic, 2007).

---

<sup>10</sup> See among others, [www.interactivebrokers.com](http://www.interactivebrokers.com) and [www.fxall.com](http://www.fxall.com)

### 2.3.2 Trading Application

To evaluate the performance of the technical trading rules two metrics for adjusted returns are used. The first measure is annualized excess return that accounts for the returns relative to the risk-free rate benchmark. The second one is the Sharpe ratio that corresponds to the risk-adjusted returns. The Sharpe ratio is the most common performance evaluation used by asset managers (Rime *et al.*, 2010). The equal weight<sup>11</sup> trading performance of the identified genuine technical rules is presented below.

#### [Table 2.2]

From Table 2.2, I note that in all cases the data snooping procedure was able to identify genuinely profitable trading rules based on the IS observations. The number of significant rules corresponds roughly to 5%-15% of the total number of trading rules under study. The average trading performance of the genuine trading rules is positive in all IS cases and in some OOS cases. The patterns are the same between the annualized excess return and the Sharpe ratios.

These results allow me to argue that technical analysis seems to have value on the exchange rates and periods under study. There are genuine profitable simple technical rules in the IS. It is possible for investors and researchers to identify these rules with the help of recent developments in SI. This performance supports AMH, which argues that investors need to be adaptive in highly competitive trading environments. These results agree with Hsu *et al.* (2010), Neely and Weller (2013) and Hsu *et al.* (2014) that argue that technical analysis has some value. However, in line with the previous studies, I note that the performance of these rules is volatile probably due to the time-varying market

---

<sup>11</sup> The equal weight corresponds to investing the  $1/C$  of the total wealth to each of the  $C$  trading rules identified from the data snooping procedure. The portfolio construction approach as presented in Bajgrowicz and Scallet (2012) has also been explored. In Bajgrowicz and Scallet (2012), the buy and sell signals counter each other while the neutral signs are considered risk free investments. The portfolios derived from this approach do not change the view of Table 2.1. However, as the scope of this study is to check the efficiency of technical analysis in FX and whether Bayesian techniques can improve their trading performance, the annualized averages are presented. Following this approach, the results of Table 2.2 can also be compared with the results of Section 2.3.3 where the best trading rules are combined with the Bayesian techniques.

conditions. I also note that the profit margins are low, and the OOS trading performance is not always above the risk-free rate<sup>12</sup>.

### 2.3.3 Bayesian Methods

Although technical rules seem unreliable for trading, Bayesian techniques can offer an advantage to investors. Arguably, they could combine different trading signals and derive strongly positive trading performances. Complex models should be capable of encompassing the simple trading rules. Additionally, these models should be able to offer an advantage to highly competitive markets. More specifically, the dynamic nature of DMS and DMA and the non-linear adaptive nature of BNN should be able to handle the changing trends of the FX series under study. However, all these three methods are computationally demanding and combining them with all the identified trading rules is not feasible<sup>13</sup>. Thus, for the DMA, DMS and BNN the best 5, 10 and 15 technical rules are used as inputs based on accuracy, profitability and Sharpe ratio in the IS. For RVM, the algorithm requires a large set of potential predictors in order to identify the optimum relevant subset of inputs. Therefore, it is fed with all the identified genuine profitable technical rules. In Tables 2.3 to 2.5, the trading performance of all the Bayesian methods in the OOS is presented. A Simple Average (SA) is also estimated as a naïve benchmark. The Giacomini and White (2006) test is applied to all combinations. The benchmark of the test is a simple random walk with no trend.

#### [Tables 2.3 to 2.5]

Tables 2.3 to 2.5 show that all Bayesian combinations are capable of producing positive returns and Sharpe ratios (see Appendix A.2), after transaction

---

<sup>12</sup> As risk free rate, the effective federal funds rate is considered. The interest rate at which US depository institutions trade federal funds with each other overnight.

<sup>13</sup> For example, for DMA and the EUR/USD in the first forecasting exercise, the algorithm would have to estimate  $2^{839}$  combinations. This task is feasible with the help of supercomputers (which were not available in this project) but it is unrealistic from a trading perspective where speed is essential. For an up-to-date personal computer (Intel core-i5 3470 64-bit processor with 8 GB memory), DMA needs around 30 mins to produce the results for one experiment with fifteen inputs (out of the twelve similar experiments, for each of the three exchange rates). This is increased to twenty-one hours for twenty inputs. Similarly, in BNNs, when the number of inputs is very large their algorithm becomes insufficient, they become prone to overfitting and their forecasting performance is crippled (Zhang *et al.*, 1998; and Yegnanarayana, 2009).

costs for the exchange rates and the periods under study. The DMA and the BNN seem to outperform their Bayesian counterparts and the SAs. This can be explained by the dynamic nature and time-varying coefficients of DMA and the highly non-linear features of BNN. On the other hand, the RVM that explores the whole set of genuine rules presents a marginally better performance than NB. DMS presents a consistent lower trading performance than the relevant DMA models. In trading, model averaging almost always works better than model selection and thus these results are not surprising. In general, I find that the profitability increases three to four times with DMA and BNN, compared to the pool of surviving technical rules (see Table 2.2). In terms of risk, the Sharpe ratios for DMA and BNN are consistently positive with an average of 0.6<sup>14</sup> for both models with 15 inputs. The same metric is 0.2 for the OOS in Table 2.2. The comparison of the Sharpe ratio for the Bayesian models and survivors of the data snooping test shows major improvement in trading performance after adjusting for risk. It is also worth noting that all Bayesian combinations are statistically different from a random walk forecasts based on the Giacomini and White (2006) test.

Based on these results, Bayesian Statistics has value in trading and can considerably increase the profitability of the underlying trading systems. The models under study (DMA, DMS, BNN and RVM) are characterized by their complexity but can offer investors substantially increased returns. Similar to the concept of AMH, in highly competitive markets (such as FX) simple rules have a small value. Traders should seek complex non-linear models that can offer them an advantage over their competitors. DMA and DMS search all possible input combinations and select the optimal subset at each step, while BNN imitates the work of biological neurons and maps the non-linear dataset through Bayesian statistics. Unlike the simple technical rules that can be estimated by hand, none of the three Bayesian models can be used without the help of computer processors. However, complexity is always translated to an increased computational burden. This study was limited to subsets of the genuine profitable

---

<sup>14</sup> The trading performance is directly dependant on the modelling process, the empirical design, and the study period of a research. Thus, a direct comparison of the trading models from different studies directly may not be accurate unless similar conditions are replicated. However, the reported value for average Sharpe ratio is comparable to 0.4 in Neely *et al.* (2009) as a standard level in OOS for FX and 0.5 for the leading models in Neely and Weller (2013).

technical rules for DMA, DMS and BNN. While this protects the models from overfitting, applying the whole set of genuine rules might have led to better results.

## 2.4 Conclusions

In this study, I explore the utility of technical analysis and Bayesian Statistics in trading. For this purpose, 7846 technical rules are generated for the EUR/USD, GBP/USD and the USD/JPY exchange rates. Then, the genuinely profitable trading rules are identified with the help of the Romano *et al.* (2008) data snooping test combined with the balancing procedure of Romano and Wolf (2010). Finally, the profitable rules are combined with NB, RVM, DMA, DMS and BNN models. The motivation for this research is the AMH which states that complex models should have an advantage in highly competitive markets. The promising forecasting performance of Bayesian Statistics in this Chapter's study confirms the proposals of the AMH and validates the paradigm of Figure 1.1.

In the results, I find that this Chapter's data snooping procedure identifies 5% to 15% of the technical rules as genuinely profitable. However, the generated portfolios based on them, present small annualized returns and Sharpe ratios over the OOS. When subsets of these rules are combined with the Bayesian models, I find that all Bayesian techniques increase the trading performance of the simple technical rules up to four times. Among the competing models, the DMA and the BNN clearly outperform their benchmarks. These results allow me to argue that market efficiency is variable, and it is possible to benefit from market inefficiencies with Bayesian Statistics.

This Chapter's results should go forward to convince traders and academics, to explore the recent development in statistics for procedures capable of providing an advantage in financial markets. These procedures might be characterized by complexity and are therefore inappropriate for high-frequency trading or large experiments. Nevertheless, the complex procedures in this Chapter can provide an edge in comparison to the traditional trading models.

## 3. Conditional Fuzzy Inference: Applications in Football Results Forecasting

### 3.1 Introduction

This Chapter introduces the concept of CF inference. In CF a set of FRs is generated in the IS and a power to each rule is assigned based on the rule's frequency and accuracy. Then, all rules are ranked and the strongest are applied in the OOS. For each OOS data point, membership functions are calculated and if certain conditions are met a weighted average of the strongest FRs is estimated. These conditions are based on the IS and ensure that the rule is strong enough for OOS estimation. If the conditions for a data point are not satisfied and no strong rule is close, then no forecast is generated for that point. The CF process avoids weak rules and ensures that the generated forecasts are based on a weighted average of the most powerful FRs.

The proposed methodology is advantageous to a series of issues. Firstly, it is useful in problems where the practitioner or the researcher is interested only in strong signals and the risk of having a poor forecast is greater than having no forecast. For example, in financial trading, betting on a sport or in any other sensitive decision-making process, poor forecasts lead to financial losses. In these environments the underlying series are volatile and decision makers are risk averse; abstaining from the market is better than making decisions under uncertainty. Secondly, CF can improve the OOS accuracy of the underlying system and offer transparency at the same time. This is beneficial in problems where complex models (such as ML) are necessary. Thirdly, the generated rules can be easily applied by non-experts as the number of rules applied is small and they are easily replicated. Lastly, the chosen rules do not suffer from over-fitting or under-fitting problems. In ML and complex models, it is common for the performance to be driven by extensive experimentation. In these cases, the generated forecasts can be due to over-specification and have no generalisation value. In CF the weak signals are dropped and the noise within the model is reduced.

A novel empirical study is designed and implemented in order to test the merits of the proposed methodology. CF and RVM will be applied to the most popular forecasting exercise in Europe and Asia, namely betting on European

football<sup>15</sup> games. Football games forecasting is a high return/risk exercise; a correct bet can offer substantial profits while a wrong decision leads usually to the total loss of the capital. Thus, it seems perfect exercise for the CF approach. More specifically, the proposed model forecasts both the result and the number of goals within football games at the three biggest football championships (the English Premier League, Italian Seria A and Spanish La Liga) from 2005 to 2016. The forecasts are evaluated through a realistic betting exercise based on the Betbrain average odds<sup>16</sup>. The aim of the exercise is to extract CF rules from selected features by RVM that are easily interpretable and can offer substantial profits to those who gamble on football games. The CF and RVM forecasts are benchmarked by those generated by an RVM model combined with an ANFIS model (the most popular fuzzy extraction structure), those from a single RVM model and those from an OP model. The comparison between CF and ANFIS will reveal the benefits of the proposed framework compared to an unconditional (ordinary) Fuzzy Inference Systems (FIS), while by benchmarking the results with the single RVM it will validate whether CF can increase the accuracy of the underlying system. OP is the most popular technique in football games forecasting. Thus, the comparison with CF offers a benchmark close to the relevant literature. In addition to the above, alternative football betting types are explored. There are three main forms of gambling in football games: betting on the result of the game (home win-draw-away win), betting on the result through the Asian handicap (win-lose) and betting on the number of goals of a game (over or under 2.5 goals). The exercise will reveal whether consistent profits from football betting are possible and if the size of the profits differs between the three forms or the football championship.

FL is introduced by Zadeh (1965). Its motivation is driven by the work and functioning of the human mind. Even though a tremendous amount of information presents itself to a human in any given situation – an amount that would ‘choke’ a typical computer – the human mind has the ability to discard the irrelevant elements and to concentrate only on the information that is relevant. The ability of the human mind to deal only with the part of the information that is relevant

---

<sup>15</sup> Simply football for the remaining of the Chapter.

<sup>16</sup> Betbrain collects odds from 138 bookmakers and betting exchanges from across the globe.

is connected with its possibility to process fuzzy information (Zadeh, 1983). In that way the amount of information the brain has to deal with is reduced to a manageable level and the decision process is faster, simpler and more effective. FL can be described as multi-value logic that allows for intermediate values instead of the conventional evaluations like yes/no, up/low, on/off. These linguistic expressions are called FRs. FRs are conditional statements in the “*if ... then ...*” form. They are widely used in studies with systems whose actions are incomprehensible (Bellman and Zadeh, 1970). In these studies, researchers apply an FIS in the IS to generate a set of FRs that maps the interactions between variables (see among others, Hruschka, 1988; Teodorović, 1994; Piramuthu, 1999; Kuo, 2001; and Chang *et al.*, 2008; and Gradojevic and Gençay; 2013). This set of FRs reveals the structure of the system and how inputs relate to the outputs. However, this structure is rarely retained to the same extent in the OOS. In problems with dynamic data (such as finance or economic series or football games results) the relationship between inputs-outputs varies through time and the generated FRs in the IS cover only a part of the OOS. This leads to a reduction of accuracy and to systems that are interpretable but inaccurate. Another problem is the strength of the generated IS FRs. Different rules have different strengths and different degrees of accuracy. Applying weak FRs in the OOS leads to poor forecasts.

There are a few studies that apply FL to the context of sports forecasting. Rotshtein *et al.* (2005) propose a model for predicting the result of a football match from the previous results of both teams. Although they suggest that their model accounts for nonlinear dependencies through fuzzy knowledge, they are focusing on a very illiquid football betting market, the one of Finland, and they do not explore the betting profitability of their forecasts. Trawinski (2010) propose a fuzzy model for extracting FRs in order to predict basketball game results. The author compares ten FR learning algorithms against a standard OLS, but they do not present robust empirical results or any betting application of football. Finally, Bastos *et al.* (2013) propose a static and a dynamic Poisson-Gamma model to predict the outcome of World Cup football results based on the number of goals scored by each team. In their application, a fuzzy C-means algorithm is used for clustering. Nonetheless, they do not offer any betting



application, while their forecasting exercise is limited to a football event occurring only every four years.

The majority of researchers in football games forecasting apply OP, probabilistic or SVM methods. OP applications are common in the literature of football forecasting due to the ordered nature of the football result variable, as a result can be ordered as away team winning, draw and home team winning the game. Kuypers (2000) applies OP in order to test how the betting market participants utilize available information and claims that an expected profit maximizing bookmaker could set market inefficient odds. That suggests that betting arbitrage is indeed possible. Audas *et al.* (2002) use OP to forecast games outcomes in English football and examine the effect of managerial change on teams' performances. One of their main findings is that within-season managerial change could be attributed to the fact that owners are willing to gamble in order to stave off the threat of relegation. Dobson and Goddard (2003) test the persistence of sequences in match results through OP and Monte Carlo analysis. Their study utilizes a 30-season English Football League and Premier League dataset and they find that there is negative persistence for sequences of consecutive wins and sequences of consecutive matches without a win. Goddard and Asimakopoulos (2004) and Goddard (2005) both apply OP to forecast English football outcomes based on teams' quality and past performance indicators. Their results indicate that this approach – which is followed in this Chapter – is robust and provides high forecasting performance. The work of Forrest *et al.* (2005) investigates whether odd-setters can forecast implicitly English football results. They use as a benchmark an OP model and they find that bookmakers forecasts are improving over time, while OP fails to outperform them. Graham and Stott (2008) show that there are systematic biases in bookmakers' odds, but they do not manage to achieve betting profitability applying OP forecasts. Finally, Forrest and Simmons (2008) examine the efficiency of betting odds in the Spanish La Liga. They argue that betting odds are influenced by the relative number of fans of each club in a match, with supporters of the more popular teams being offered higher odds.

Bayesian and Poisson probabilistic approaches are also found in this strand of the literature. Meeden (1981) demonstrates the optimal strategy for a Bayesian

bettor playing against a Bayesian bookmaker. Dixon and Coles (1997) forecast the number of goals and the results of football games with forecasting systems based on the Poisson distribution. Rue and Salvesen (2000) apply successfully Bayesian Statistics and Markov chain Monte Carlo iterative simulation techniques to the task of forecasting the final ranking of Premier League and how each team's properties vary during the season 1997-1998. Crowder *et al.* (2002) estimate the probabilities of home win, draw and away win with refinements of the independent Poisson model for 92 soccer teams in the English Football Association fixtures over 1992-1997. A Bayesian Network (BN) is used by Joseph *et al.* (2006) to predict football forecasting results. Although their approach provides high statistical accuracy, it is only focusing only on one English Premier League team. Vlastakis *et al.* (2009) applies Poisson count and multinomial logit regressions in order to encompass forecasts for football betting. Their results suggest that there is evidence of weak-form efficiency in the betting market. The work of Karlis and Ntzoufras (2008) proposes a Bayesian approach to model the goal margins, while Min *et al.* (2008) combine Bayesian inference and rule-based reasoning in order to forecast the result of football games. Baio and Blangiardo (2010) propose a Bayesian Hierarchical model to predict the results of the Italian Serie A based on the defensive/offensive mentality of the team. Constantinou *et al.* (2012) suggest a pi-BN for English Premier league match forecasting. Their model applies time-dependent data with weighted degrees of uncertainty and exhibits high statistical accuracy along with profitability over the publicly available odds. A similar approach is applied by Owramipur *et al.* (2013). Their proposed BN is applied to the task of forecasting Barcelona FC's results in the 2008-2009 Spanish league using as inputs psychological and non-psychological factors that can affect the team's performance. Finally, Koopman and Lit (2015) and Angelini and De Angelis (2017) refine previously traditional Poisson models and improve upon the accuracy of the Dixon and Coles (1997) model. Their applications are based on the Premier league and their profitability is tested with a basic betting strategy.

Other studies in the field utilize SVMs frameworks. Vlastakis *et al.* (2008) applies SVMs to the task of predicting power over European football match scores using match result and Asian Handicap odds data from the English Premier league. SVMs performs better than a Poisson model and the authors note that the size of the Asian Handicap appears to be a significant predictor of both home and away

team scores. Gomes *et al.* (2016) propose a pervasive decision support system that utilizes SVM to predict the number of football corners and goals of Premier League, while in a similar application Martins *et al.* (2017) consider SVM as benchmark for their football match results' forecasts of several championships. Recently, Baboota and Kaur (2018) compare the forecasting performance of several techniques on predicting the rank probability scores for the English Premier League using teams' past performance indicators as inputs. Their results suggest that linear SVM is not performing well, but SVM with radial basis kernel efficiently predicts the home and away wins outperforming other techniques such as Naïve Bayes.

With the previous background in mind, the following table positions this Chapter's proposed methodology in the above literature and illustrates how this work differs from other relevant studies. None of the previous studies, apply a conditional FL framework similar to the one introduced in this Chapter or has extensive betting application that incorporates the Kelly criterion.

### [Table 3.1]

In a nutshell, Operational Research (OR) methods seem promising in football metrics. The recent literature argues that the football betting market is at least weak efficient (see among others, Kuypers, 2000; Goddard and Asimakopoulos, 2004; and Vlastakis *et al.*, 2009). The above literature makes football games forecasting an ideal fieldwork for the proposed CF. It is a forecasting exercise that is characterized by its difficulty, uncertainty and high public interest.

The remainder of this Chapter is organised as follows. Section 3.2 introduces the concept of CF and describes in detail the methodology. Section 3.3 describes the dataset used and the empirical application on football betting. In Section 3.4 some concluding remarks are presented, while in Appendix B several technical details are explored and the mathematical background on ANFIS and OP are provided.

## 3.2 Methodology

This Section summarizes all the relevant information regarding the proposed methodology. Initially, the CF method is motivated. Then, its mathematical and technical description is explained in detail. Finally, short descriptions of the benchmarks and Kelly criterion are provided.

### 3.2.1 RVM – Underlying System

RVM is a sparse kernel model to tackle the problem of large-scale data-processing. The Bayesian learning framework used in RVM can generate precise forecasts while reducing the feature space to the most important (“relevant”) vector. RVM’s solid probabilistic approach allows the inference of the optimal hyperparameters and vectors’ weights from the data. In this Chapter’s proposed approach, RVM will serve for both feature selection and prediction. The selected features (relevance vectors) are applied to an ANFIS and a novel extension of FIS named CF. For the mathematical details of ANFIS and RVM, the reader is referred to Appendix B.4 and Section 2.2.2 respectively.

### 3.2.2 Conditional Fuzzy Inference

#### 3.2.2.1 Motives

FL frameworks are traditionally applied in decision processes, where the level of uncertainty is high and a complete problem formulation is difficult. In FL, an FIS is applied and a set of FRs are extracted for prediction and decision-making (Sugeno, 1985)<sup>17</sup>. The most popular FIS is the ANFIS of Jang (1993).

ANFIS is an FIS framework that utilizes the benefits of ANNs and FL principles (Polat and Güneş, 2007). The training algorithm of ANFIS is forward and backward oriented, which is a common approach in OR applications<sup>18</sup>. Nonetheless, the

---

<sup>17</sup> For the identification and parameterization of a FIS the interested reader should refer to studies such as Gustafson and Kessel (1979), Bezdek *et al.* (1984), Jang (1993), Chiu (1994) and Angelov and Filev (2004).

<sup>18</sup> In the forward stage, the premise parameters are fixed, and the consequent parameters are estimated by the least squares method. In the backwards one, the consequent parameters are kept fixed and the errors are back-propagated. Then, the premise parameters are optimised through the gradient descent method (Shapiro, 2002).

methods of defuzzification vary in the literature. For example, Jang (1993) classifies possible defuzzification functions in three main categories. In the first category a crisp output for each rule is estimated and then a weighted average of each rule's output forms the aggregate output as the fuzzy model's output. The applied weights are based on the firing strength of the rules and the output membership function. The second category of defuzzifiers involves applying a "max" operator to qualified fuzzy outputs that meet some criteria (e.g. minimum firing strength) and then the aggregate output is determined by a function such as the mean of maxima, the maximum criterion and so forth. The third category is the Sugeno approach, namely for each rule's output, a general linear model based on inputs is estimated and the aggregate output is the weighted average of all rules involved. The ANFIS model follows the third type of defuzzification method. For the exact mathematical formulation of ANFIS, I refer the reader to Jang (1993).

The conditional approach in the second category that is overlooked in the ANFIS structure can be promising. If all FRs extracted through the training process are not of the same quality, the user might prefer to use the FRs partially rather than entirely. The ANFIS adopts the double pass algorithm to optimise the rule specifications but may lead to over-fitting and to undesirable performance for the previously unobserved points. In other words, the problem is that in the OOS the membership grade may not be strong enough for some or all rules. This eventually leads to bad forecasts. This problem is common in ML where the OOS performance is either considerably worse than the IS's one (over-fitting), or the model is inadequately trained (under-fitting).

The proposed methodology in this study combines the merits of the second and the third defuzzification approach. A crisp output based on a simple linear model is highly favourable in case of large-scale problems. On the other hand, when a dataset comes with remarkable noise a selective procedure for the FRs aggregation can control the uncertainty. The next Section explains the algorithm for the CF.

### 3.2.2.2 Algorithm

Let me define the firing strength for rule  $i = 1, \dots, M$  as:

$$\mathcal{W}_i^* = \prod_{k=1}^K \dot{\mu}_{ik}(x_k^*, mf) \quad (3.1)$$

where  $\dot{\mu}(\cdot)$  is the membership grade calculated based on a given membership function  $mf$  and a  $K$ -dimensional test data point with input vector  $x^* = [x_1^*, \dots, x_K^*]$ .

The choice of the membership function can be central to the model's fit and its predictive ability. Based on Guillaume (2001), for most inputs of a fuzzy system, it can be intuitively accepted that as an observation places farther from the centre of a rule, the associated membership grade falls. Thus, it can be assumed that the membership function needs to come with a bell shape. The Gaussian membership function offers such smooth shape around centres of clusters and therefore it is a fitting one for this study. Such choice is suitable for forecasting applications like this Chapter's (see among others, Wang and Mendel, 1992; Jang, 1993; Kunchewa, 2000; Cheng and Lee, 2001; and Akkoç, 2012). For football forecasting specifically, Vlastakis *et al.* (2008), Igiri (2015) and Baboota and Kaur (2018) that apply SVM to football match prediction, employ the Gaussian kernel to their models. Other membership functions (i.e. univariate/multivariate sigmoid, triangular, and trapezoid) were also experimented in the IS. In all cases, the IS accuracy was worse than the one acquired with the Gaussian membership function. Under a Gaussian distribution assumption, the membership grade for the  $k$ -th element of rule  $i$  is given by:

$$\dot{\mu}_{ik}(x_k^*, mf) = \exp \left\{ - \left( \frac{x_k^* - c_{ik}}{a_{ik}} \right)^2 \right\} \quad (3.2)$$

where  $a_{ik}$  and  $c_{ik}$  are specified for the FRs through the training process.

The CF proposes eligibility criteria for the defuzzifier function to discard rules with unsatisfactory membership grades from the set of rules that form the fuzzy output. This process involves two components: the criteria set up and a satisfactory threshold ( $\Theta$ ). The criteria set up determines the evaluation function for each rule and can take one of the forms offered in Eq.s (3.3 and 3.4) below:

$$C_i^* = \begin{cases} 1, & \mathcal{W}_i^{*1/K} \geq \Theta \\ 0, & \text{Otherwise} \end{cases} \quad (3.3)$$

where  $\mathcal{W}_{x^*,i}^{1/K}$  is given by Eq. (3.1). Or alternatively:

$$C_i^* = \begin{cases} 1, & \min_k(\mu_{ik}(x_k^*, mf)) \geq \Theta \\ 0, & \text{Otherwise} \end{cases} \quad (3.4)$$

The criteria set up can be defined as an average of membership grades for each rule compared to  $\Theta$  (Eq. (3.3)) or as the minimum membership grade compared with  $\Theta$  (Eq. (3.4)). Both procedures generate an indicator whether the rule should be applied or not. In the Eq.s (3.3 and 3.4),  $C_i^*$  is the indicator to determine whether the rule  $i$  is qualified ( $C_i^* = 1$ ) as an eligible rule for the specific test observation to be applied. Alternatively, the rule is considered as weak if  $C_i^* = 0$ . The other component to shape the eligibility criteria is the threshold parameter.  $\Theta$  determines the average power required for a rule to be considered as a strong rule. I propose defining the  $\Theta$  as:

$$\Theta = \max(w_\lambda^{1/K}, \Lambda) \quad (3.5)$$

Under this setting, the threshold level originates from two sources. Firstly, from the training dataset reflected in  $w_\lambda^{1/K}$  and secondly, from a general value allocated to membership grade represent a strong rule ( $\Lambda$ ).  $\Theta$  is simply the larger of the two quantities. Alternative settings for  $\Theta$  depending on the nature of application might be used as well. In Eq. (3.5)  $w_\lambda^{1/K}$  is the  $K$ -th root of average firing strength of a rule  $\lambda$  that is the minimum acceptable level of firing strength based on the training dataset. In other words,  $w_\lambda^{1/K}$  is the endogenous threshold based on the IS. To set  $\lambda$  the following procedure is pursued:

Firstly, the arithmetic average membership grade for each rule  $i$  and input element  $k$  is computed over the IS dataset  $\bar{\mu}_{ik}(\cdot)$ . Secondly, the average firing strength  $w_i$  is given by:

$$w_i = \prod_{k=1}^K \bar{\mu}_{ik}(\cdot) \quad (3.6)$$

Thirdly, the FRs are sorted in a descending order of average firing strength (sorted list). In the sorted list, the rules on the top are ones that data-points place closest to the centres of the FRs. An arbitrary percentile level for the top strongest rules is selected based on the complexity of the problem and the aims of the researcher (e.g. ten). Then, the 90<sup>th</sup> percentile rule on the sorted list is selected. The equivalent index (row) of the selected rule on the original list of rules is the  $\lambda$ .

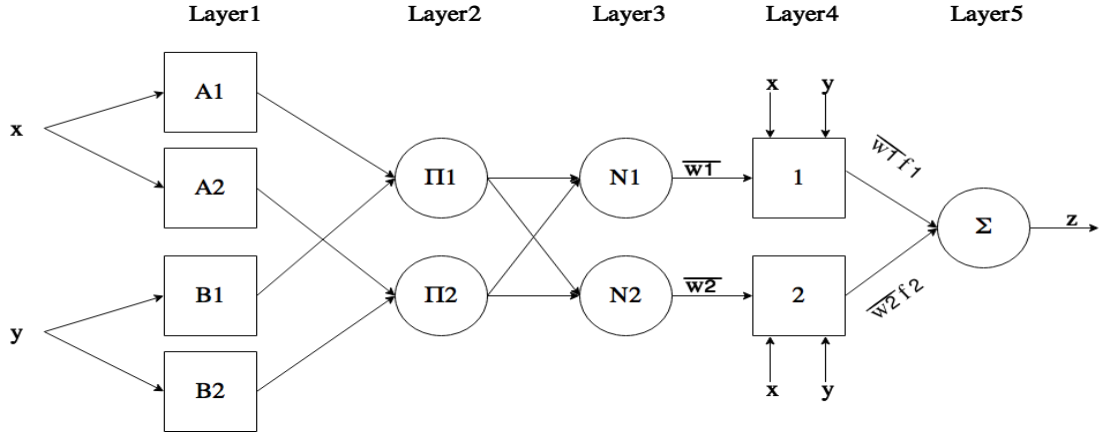
Unlike  $\lambda$  and  $w_\lambda^{1/K}$  that are endogenous,  $\Lambda$  is a fixed quantity that determines the general quality of a strong rule. The membership grade ranges from zero to one. The closer the data-point gets to the centre of a rule, the higher will be the membership grade. Ideally, if the data-point and the rule centre overlap, the membership grade is equal to one. The further the data-point gets, the lower the membership grade becomes. This implies that the FR is getting weaker. The choice of  $\Lambda$  depends on the problem under study and the practitioners' needs. In problems where uncertainty is high, and a wrong forecast can have a considerable impact,  $\Lambda$  should be set high (over 0.8). This will reduce on the one hand the uncertainty but on the other hand, it will also reduce the number of CF forecasts generated. In problems where wrong forecasts have a small effect on the utility function of the practitioner,  $\Lambda$  can be set lower. This will lead to more CF forecasts that retain some level of risk.

The combination of the endogenous and the exogenous threshold ensures that the applied rules for forecasting are correctly fitted. Over-fitting and under-fitting are the most significant challenges in ML. If an FIS is over-fitted, the average membership grades, the average firing strengths and accordingly the  $w_\lambda^{1/K}$  are high. As the model is overly specified to match the training samples, once the OOS data-points are fed the average membership grades for the new observations will fall below the  $w_\lambda^{1/K}$ . In the CF, the model will drop these rules and looks for any remaining rule that can satisfy the  $w_\lambda^{1/K}$  threshold. This ensures that model forecasts properly even if the original FIS is over-fit. If the CF is unable to identify for a single point any relevant strong rule, then no forecast is generated. Similarly, in case of an under-fit model, the IS measure  $w_\lambda^{1/K}$  is low. However, there might be still some rules that are strong and fit enough for certain points in the OOS.  $\Lambda$  ensures that under-fit rules with low  $w_\lambda^{1/K}$  are not applied.

In order to grasp the main contribution of the CF over the ANFIS, the five-layer ANFIS structure is presented in Figure 3.1 below:



Figure 3.1: ANFIS Architecture



**Note:** This is a typical five-layer ANFIS structure, assuming A and B fuzzy sets,  $f_1(x, y)$  and  $f_2(x, y)$  the estimated outcomes for each rule through a polynomial combination of the inputs and  $O^1, O^2, O^3, O^4, O^5$  the outputs of each layer. First layer stores the parameters of cluster centres for each input and the membership grade for each input is calculated. In the second layer, the membership grades are aggregated for each rule through the  $\Pi$  operator. The outcome of this layer is the firing strength representing the power of the associated rule ( $w_i = \mu_{A_i}(x) \times \mu_{B_i}(y)$ ,  $i = 1, 2$ ). In the third layer, the firing strengths are adjusted by the normalizing operators (N1 and N2) for each fuzzy rule. Layer 4 nodes are adaptive as they are being fed with inputs  $x, y$  to generate the output for each fuzzy rule separately  $\bar{w}_i f_i(x, y)$ ,  $i = 1, 2$ . Finally, the defuzzified ANFIS realizations in the last node are a simple aggregation of its inputs via sum operator  $\Sigma$ .

The proposed CF modifies the last layer. The modification is able to grasp the strongest possible FRs and drop the mediocre and poor ones. In the CF model, the final layer outcome can be presented as:

$$\hat{O}^* = \hat{O}_5(x^*, O_4, O_3, O_2, O_1, A, C, P, Q, R, \lambda, \Lambda) \quad (3.7)$$

As the equation implies the difference in CF compared to ANFIS is in the defuzzifier module, where the aggregate output is modified to include the eligibility criteria. The CF output  $\hat{O}^*$  is computed through a node function  $\hat{O}_5$  given the input vector  $x^*$ , ANFIS specification, and threshold parameters. The ANFIS specification include training nodes ( $O_1, \dots, O_4$ ), premise parameters set  $\{A, C\}$  and consequent parameters set  $\{P, Q, R\}$ . The CF threshold parameters are  $\{\lambda, \Lambda\}$ . I propose two defuzzification functions. Firstly, by calculating a weighted average of outputs for qualified FRs and secondly the output of the strongest qualified rule represented in Eq.s (3.8 and 3.9) below:

$$O'_5 = \frac{\sum_{i=1}^M w_i C_i f_i}{\sum_{i=1}^M w_i C_i} \quad (3.8)$$

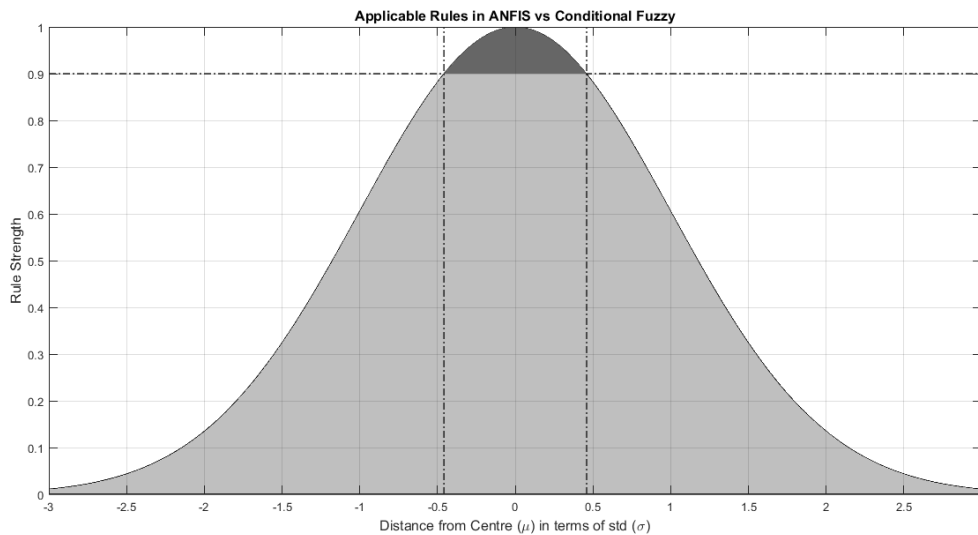
Or alternatively:

$$O'_5 = f_v(.), v = \arg \max_i (\omega_i) \quad (3.9)$$

In Eq.s (3.8 and 3.9),  $O'_5$  is the node function for the last layer of CF. For each rule  $i$  the firing strength  $\omega_i$  is estimated by Eq. (3.1), while the conditional indicator  $C_i$  is given by Eq. (3.3 or 3.4) based on threshold parameters estimated in Eq.s (3.5 and 3.6).  $f_i$  is the regression output estimated by the FR (traditionally as with any Sugeno fuzzy approach). Finally, in Eq. (3.9)  $v$  is the argument of maxima for the firing strength  $\omega_i$ .

The above-mentioned modification is very crucial, as it can provide predictions for points (observations) where strong rules are nearby and at the same time satisfy endogenous and exogenous threshold specifications. The final CF outcome is a weighted average only of the strongest rules. This attribute is innovative in FL, as it is able to offer interpretability of the final result, protection against substantial forecast errors and under- or over-fitting in the underlying decision-making system. To enlighten the novelty of the CF method, a comparative graph of the rules selected in the case of CF and ANFIS is presented in Figure 3.2.

**Figure 3.2: Rule selection comparison between CF and ANFIS**

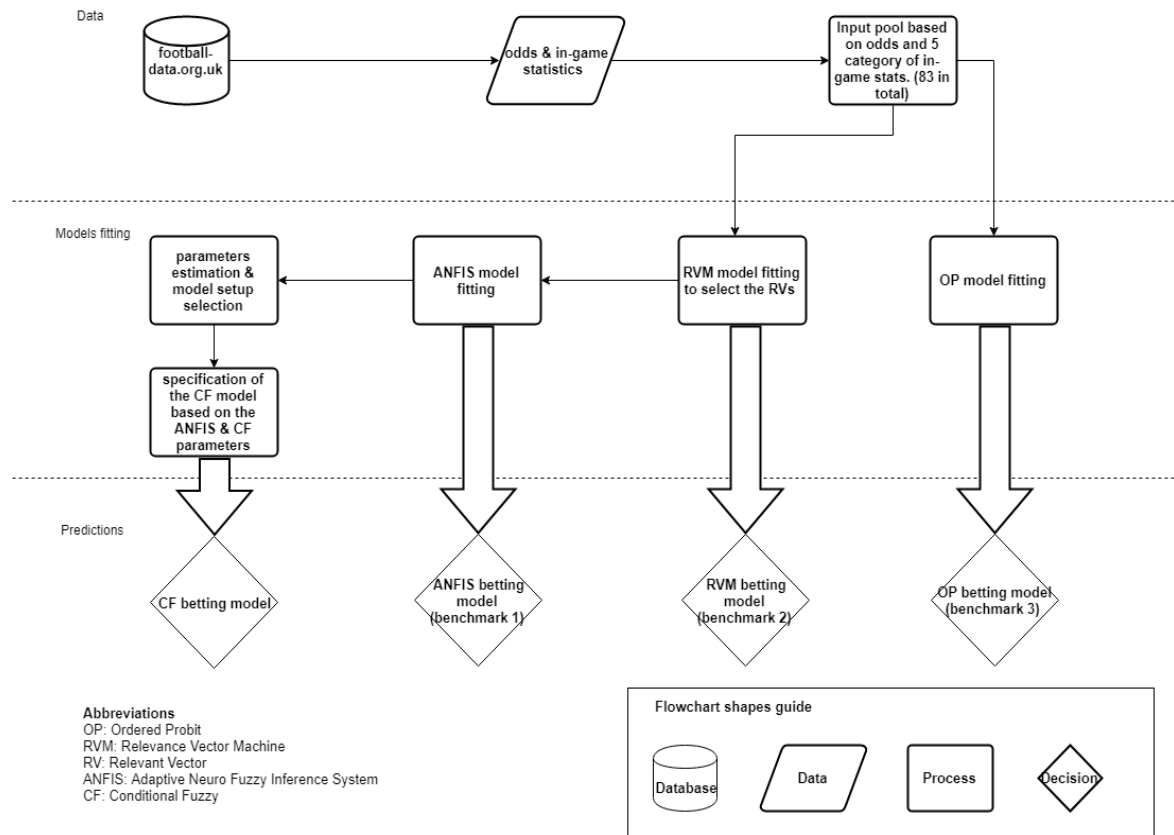


**Note:** The figure shows the selection of fuzzy rules with respect to the distance from centres ( $\mu$ ) in units of standard deviation ( $\sigma$ ) and the corresponding strength. In ANFIS (whole distribution) all fuzzy rules are accepted irrespective of the strength, but in CF (dark grey peak of the distribution) only the strongest rules are selected to for defuzzification stage.

From the figure above, it is clear that CF is selecting rules that are on the peak of the distribution of the ANFIS selection and within small distances from the

centre. The flowchart for the modelling module of the Chapter is presented in Figure 3.3. The flowchart shows the steps taken for the model synthesis of CF.

**Figure 3.3: Modelling Flowchart of Chapter 3.**



**Note:** The study utilizes the generated input pool to generate four sets of predictions. The proposed methodology is the CF betting model. In the modelling process of this study, the OP model is fed directly the input pool to generate forecasts. The RVM is also fed the whole input pool and generates the RVs. Based on the RVs, its performance is explored, while the selected RVs are the inputs of the ANFIS and CF.

Figure 3.3 shows that in the first stage, the system inputs are fed into RVM that selects the RVs and derives a series of forecasts. Then based on these forecasts, the CF approach is applied and a series of FRs, their associated membership function and their firing strength are generated. In this Chapter's application, I apply CF following the evaluation function of Eq. (3.3)<sup>19</sup>. Here it should be noted that the CF model synthesis is based upon the algebra offered by Zadeh (1965), Jang (1993) and Singpurwalla and Booker (2004). The alternative definition of the indicator in Eq.s (3.3 and 3.4) are based on AND and PRODUCT operators in the FL context, whereas the defuzzification methods in Eq.s (3.8 and 3.9) are based on the

<sup>19</sup> The results of my empirical application are almost similar with both evaluation functions (Eq. (3.3 or 3.4)) and defuzzification functions (Eq. (3.8 or 3.9)).

discussion of Jang (1993). The theoretical analysis of this algebra is provided in Singpurwalla and Booker (2004).

In terms of applying CF, football betting is a high risk-high return exercise. A wrong bet usually results in the total loss of the invested capital. Thus,  $\Lambda$  is set high to 0.90. This will result in a stricter selection for CF and resembles the real-world practice where bettors are highly selective on the games that they will bet. In the OOS, for each point the membership grades for the FRs are estimated and based on the eligibility criteria, a signal might be generated. In case a signal is generated the relevant bet is placed. If a signal is not generated, I abstain from the market. A detailed illustrative example of the CF application is provided in Appendix B.1.

### 3.2.3 Benchmarks

This approach is benchmarked as mentioned earlier against a basket of forecasting models, namely the OP, the RVM solely and the ANFIS using the selected RVs from the RVM. The flowchart of Figure 3.3 makes apparent that the CF model needs to be benchmarked with ANFIS. With this comparison, I will be able to quantify if the modification of the last layer of ANFIS is successful and whether the conditional approach is beneficial.

The benchmark selection can establish the gradual performance change among candidate models (RVM, ANFIS, CF and OP). RVM can optimize the global parameters that affect the input variable space. Its Bayesian probabilistic approach is beneficial by firstly producing sparse solutions able to reduce the input space for other models and secondly optimal parameters that allow the RVM to forecast efficiently. This is very important, as traditional methods such as cross-validation are not able to achieve this (Tipping, 2001). Therefore, RVM is used as a screening procedure for generating the reduced input set for ANFIS and CF. OP originates from Logistic Regression (LR) models and tries to estimate the probability of each outcome for a dependent variable. Depending on the number and nature of the possible choices that the dependent variable can take, the choice of the LR type can be binomial, multinomial and ordinal. Given that the outcomes of football matches can be ordered as home team winning, draw and away team winning the game, the OP model is suitable for this football betting

application. Additionally, in the literature of football betting, OP is one of the most popular methods (see among others, Koning, 2000; Forrest *et al.*, 2005; and Graham and Stott, 2008). Therefore, the use of OP model as another benchmark is justified, especially since it uses the whole input pool as in the case of RVM. The mathematical details of OP (as earlier with ANFIS) are provided in the Appendix B.5.

### 3.2.4 Kelly Criterion Application

At any investment (such as in football betting) there are three main areas to cover. These are the investment strategy, the timing of the investment (whether to invest or not), and the size of the invested capital. The investment strategy can be guided by statistical models (RVM, ANFIS, CF, and OP in this Chapter's case). The timing of the investment can be guided also by a conditional procedure like CF or the investor's preference. The optimal size of the investment or football bet can be determined through the Kelly criterion.

Consider the case of having an initial capital  $\mathcal{X}_0$ . The goal is to maximise the expected value of capital after  $n$  trial ( $\mathcal{X}_n$ ). Now suppose that a gambler is interested in a bet with win (loss) probability  $p$  ( $q$ ) and payoff  $b$  for every unit wager. The purpose is to maximise:

$$g(f) = E[\log(\mathcal{X}_n/\mathcal{X}_0)] = p \log(1 + bf) + q \log(1 - f) \quad (3.10)$$

where  $f$  denotes the fraction invested in the bet and  $g(f)$  is the growth based on the fraction invested in each bet. The optimal fraction ( $f^{opt}$ ) based on Kelly (1956) and Thorp (2008) is given by:

$$f^{opt} = \frac{bp - q}{b} \quad (3.11)$$

In order to apply the Kelly criterion, I need the corresponding probabilities  $p$ . These are readily available in the OP framework, as the estimated conditional probabilities of each outcome. RVM and CF have a different structure and the winning probabilities are not readily provided. However, the CF framework allows me to resemble the Kelly criterion and following a similar procedure to extract

the optimal fractions  $\beta^{opt20}$ . The firing strength can determine how probable it is to use the specific rule  $i$  for bet  $k$ . By aggregating this measure and normalising it for the rules with similar output I provide an approximated conditional probability for each ordinal outcome.

More specifically, I associate the winning probability with the conditional probability of each outcome to happen. For a bet  $k$  to be placed the normalised conditional firing strength  $\tilde{w}_i^k$  for all the eligible FRs is estimated by:

$$\tilde{w}_i^k = \frac{C_i^k \omega_i^k}{\sum_{i=1}^5 C_i^k \omega_i^k}, \forall i \quad (3.12)$$

In the next step the  $\tilde{w}_i^k$ s are aggregated for the rules that fall under same ordinal outcome:

$$\hat{\pi}_a^k = \sum_s \tilde{w}_s^k \mid f_s^k(.) \approx a \quad (3.13)$$

where the  $f_s^k(.)$  represents the predicted output of the rule  $s$ . The procedure to interpret the regression output  $f_s^k(.)$  to a class label  $a$  is the same in OP. By setting  $p = \hat{\pi}_{k,a}$  in Eq. (3.11), the optimal fraction for each outcome of the bet is given. It should be noted that when the CF does not find any eligible rule for the bet, the winning probability is zero and the  $\beta^{opt}$  becomes negative. A bet is only placed in case of having a positive  $\beta^{opt}$ .

### 3.3 Empirical Study

In this Section, the empirical application is presented. The purpose of the application is to demonstrate the merits of the CF concept and to examine whether consistent profits in football betting from simple FRs are possible. Initially, I describe the dataset used. Then, the empirical results from this Chapter's novel football betting application are presented.

---

<sup>20</sup> Singpurwalla and Booker (2004) demonstrate that probability theory has a sufficiently rich structure for incorporating fuzzy sets within its framework.

### 3.3.1 Dataset

All models are applied to forecast the results and the number of goals of football games in the English Premier League, Italian Seria A and Spanish La Liga from 2005 to 2016. The dataset of this study is publicly available at <http://www.football-data.co.uk>.

In football game result forecasting, there are three different outcomes (home win, draw and away win). An alternative approach based on the Asian handicap is also explored. Asian handicap is a form of betting on a football game result in which teams are handicapped according to their form. The strongest (favourite to win) team has a goal deficit in the start of the game while the weakest (the underdog) team has a head start. For example, assume there is a football game (team A vs team B) and team A has a handicap size of -1.5. If the game ends 2-1 and a gambler has bet on team A, they will lose the bet as the goals leads (1) is less than 1.5. Asian handicap has become increasingly popular as bookmakers offer higher winning chance compare to the traditional win-draw-lose odds. In the number of goals forecasting, gamblers bet whether the total number of goals in a game will or will not exceed 2.5.

The models under study require a series of inputs that are bookmakers' odds and past teams' performance indicators. The Betbrain average odds are used in this study. These inputs are summarized in Table 3.2.

#### [Table 3.2]

The previous literature in football forecasting applies sub-sets of Table 3.2. RVM is able to select its inputs from a large set of potential predictors through a probabilistic framework. Thus, it is not necessary to restrict the inputs to smaller sets or apply additional techniques to reduce the dimensions of the predictors' pool. For example, Dixon and Robinson (1998), Oberstone (2009), Baio and Blangiardo (2010), and Angelini and De Angelis (2017) use the number of goals scored in a match to improve forecasting accuracy of the final football outcome. Other studies consider the odds of home win/draw/away win for football game predictions (see among others, Dixon and Coles, 1997; Crowder *et al.*, 2002; Dobson and Goddard, 2003; Constantinou *et al.*, 2012; and Boshnakov *et al.*,

2017). The number of corner kicks implies offensive pressure and is considered a good proxy for higher scoring probability. Thus, Andersson *et al.* (2009) apply this variable in their football betting models. Oberstone (2011) and Martins *et al.* (2017) apply shots on target as another proxy of the offensive capacity of a team. Most of the inputs in Table 3.2 are based on the performance of the home/away team during the last three games. Incorporating teams' recent performance indicators (points of last three games, goals scored in the last three games, etc.) is crucial for the adaptiveness of the CF process and their utility are supported by Goddard and Asimakopoulos (2004), Goddard (2005), and Rotshtein *et al.* (2005). Here it should be noted that in the football betting literature it is well-accepted practice to use bookmaker odds for modelling and forecasting football outcomes (Goddard and Asimakopoulos, 2004; Štrumbelj and Šikonja, 2010; Štrumbelj, 2014; Schumaker *et al.*, 2016). Additionally, using fixed-odds, like mine from BetBrain, is considered advantageous as bettors know the final odds at the time of betting (Feess *et al.*, 2016). This is why fixed-odds betting applications are widely found in the respective literature (see among others, Dixon and Pope, 2004; Forrest and Simmons, 2008; and Constantinou *et al.*, 2012).

The implemented forecasting exercises span over the period of 2005 to 2016. The forecasts from the models are evaluated in terms of accuracy through the relevant Betbrain average odds, the average profit per bet, the proportional cumulative annualized return and the Kelly Criterion (see Section 3.2.4). The average profit per bet is defined as:

$$\frac{\sum_{q \in Q} x_q b_q - |Q|}{|Q|}, q \in Q \quad (3.14)$$

where  $Q$  is the set of games on the season that a bet is placed,  $x_q$  takes the value of 1 if the bet on game  $q$  is won based on the relevant forecast and 0 otherwise,  $b_q$  is the relevant Betbrain average odd and  $|Q|$  is the cardinality of set  $Q$ . The proportional cumulative annualized return is estimated simply by betting at each game always the 5% of the total capital which initially is 100 units. For each subsequent game, I continue to bet the 5% of the total remaining pot. The proportional cumulative annualized return is the accumulated return in the end of the season. This practice resembles the reality where gamblers bet a proportion of their wealth.



The IS consists of three football seasons<sup>21</sup> and the OOS by the following two seasons (i.e. in the first forecasting exercise for the Premiership, the football seasons 2006-2007, 2007-2008 and 2008-2009 act as IS and the seasons 2009-2010 and 2010-2011 act as OOS). The estimation is not rolled forward from the first (2009-2010) to the second season (2010-2011) of the OOS. Thus, the second OOS season act as robustness to the models. In all seasons the first three home and the first three away games of a team are discarded from the exercise. Otherwise, if team A plays against team B and one of the two teams has fewer than three home and three away games, this game is excluded from the exercise (both as IS and OOS data)<sup>22</sup>. The following Section present this Chapter's findings from all the forecasting exercises.

### 3.3.2 Empirical Results

In Tables 3.3 to 3.5 there are the accuracy ratios of all the models in the OOS while their relevant IS performance is in Appendix B.2. The number of models that CF is applied is presented in Appendix B.3. The Pesaran-Timmermann (PT) (1992) test for the aggregate performance of the models is also estimated. The null hypothesis of the test is that the relevant model is unable to classify correctly the underlying series<sup>23</sup>.

#### [Tables 3.3 to 3.5]

From the tables above, I note that the CF approach is clearly improving the accuracy of the underlying system (RVM). It generates forecasts based on the

---

<sup>21</sup> I have experimented also with two and four seasons as IS. The accuracy ratios in both IS and OOS are very close to the ones presented in Table B.4 of the Appendix B.2.

<sup>22</sup> This is happening for two reasons. It is well-known amongst football enthusiasts that the behaviour of teams at the start of the season is volatile. This happens either due to changes to the roster of the team during the summer or due to the different training during that period (for example, a team that has qualification games for the European Championships is forced to start its preparation earlier than the rest). Secondly, this process ensures that all series are smooth. For example, in the first game of the season the input series Points of H team – Points of A team would have been equal to 0 and the rest of the inputs would have to be drawn from the previous season.

<sup>23</sup> I did not estimate the PT test for each separate year for two reasons. Firstly, I am interested in the average performance of the models and not their individual accuracy for a specific year and championship. Secondly and more importantly, the CF approach selects a subsample of games and generates forecasts only for them (see Appendix B.3). In some cases, the number of CF forecasts is too small for the test and given the test's assumptions this can cripple the test's efficiency (Pesaran and Timmermann, 1992).

strongest rules which in turn leads to an improved predictability. In most of exercises, CF is up to 10% more accurate than RVM. On the other hand, ANFIS leads to slightly poorer accuracy ratios than those obtained by RVM. A finding consistent with the related literature (see Martens *et al.*, 2007). The OP presents the worst performance in game result and close accuracies to ANFIS and RVM for over-under and the Asian handicap. It is interesting to note that all models for all years, championships and seasons provide accuracies better than the ones of a random classifier (which is 33.33% for the game result and 50% for the Asian handicap and number of goals exercises). As expected, the accuracy falls in the second OOS season, but it is still higher than those of a random model.

However, the PT test reveals that only the CF is capable of classifying accurately the underlying series even two seasons ahead in all exercises. Its benchmarks seem to work well in game result but lose power in over-under and the Asian handicap exercises. I also note that the accuracies are higher for the Premiership. This might imply that the English championship has less noise or in other words is easier to be forecasted.

The previously discussed accuracy ratios might seem promising, but they do not guarantee profitability. The risk in football gambling is that almost always the bookmakers' odds differ between the seasons and championships. For example, the odds offered for the Asian handicap are lower than their game result counterparts<sup>24</sup>. Betting agencies naturally possess superior information and modify the odds by exploiting the bettor's cognitive biases in a way that mitigate their risk and increase their profitability (Cain *et al.*, 2000; Forrest and Simmons, 2008; Newall, 2017). Tables 3.6 to 3.8 present the average profit per bet for the models, championships and seasons under study.

### [Tables 3.6 to 3.8]

The CF approach seems to clearly outperform its benchmarks and offers impressive profits. The average profit per bet of CF is clearly above 10% in the first year of the OOS in all cases. This profit is substantially reduced in the second

---

<sup>24</sup> Because the number of potential outcomes in the Asian handicap (2) is less than the game result (3) the odds decrease.

year of the OOS but still remains positive. On the other hand, the other three models under study seem to present negative profits per bet in the majority of cases even at the first year of the OOS. Comparing Tables 3.6 to 3.8 with the associated accuracy results (Tables 3.3 to 3.5), I note that statistical accuracy is not synonymous of betting profitability. There are cases where the PT test indicates that the underlying model classifies accurately the underlying series and the related average profit per bet is negative.

The average profit per bet considers that the forecaster will invest in all games the same amount of capital irrespective of their previous performance. In reality, forecasters will probably modify their invested capital based on the previous record and follow a more adaptive strategy. Thus, in Tables 3.9 to 3.11 I present the proportional cumulative annualized return. This measure assumes that the forecaster always bets the 5% of their total pot. So, for example, in the first game in any give exercise the gambler will bet the 5% of their total pot which if it is 100 units which is 5 units. If the forecaster wins the bet and their earnings are 4 units, their total pot will be now 104 units. Thus, in the next game the forecaster will bet 5.2 units. The proportional cumulative annualized return offers a more realistic approach in betting where participants modify the size of their bets based on their previous record.

### [Tables 3.9 to 3.11]

I note that the pattern of the CF's profitability is similar with the one obtained by the previous metric. This profitability varies throughout the seasons and the championships but remains positive in the first year of the OOS. On the other hand, now the average profitability of CF in the second year of the OOS is not always positive. The other models present a consistent negative performance in all exercises<sup>25</sup>.

The Kelly criterion allows me to bet based on the probability of a favourable outcome. The proportion of capital that is dedicated for each game is based on the probability of winning the bet based on CF or OP. Tables 3.12 and 3.13 present

---

<sup>25</sup> The results of my empirical application are similar with both evaluation functions (Eq. (3.3 or 3.4)) and weighted average functions (Eq. (3.8 or 3.9)).

the average profit per bet of the CF and the OP based on the Kelly criterion respectively.

### [Tables 3.12 and 3.13]

From Tables 3.12 and 3.13 an increase in the profitability in most exercises of CF and OP with the Kelly criterion is observed. This increase is profound in the cases where the studied models were presenting negative results previously, namely the second OOS year of CF and all OP exercises. In first year of the OOS for CF, where CF had already positive profit per bet, the results remain similar. As discussed earlier bookmakers' odds are biased to exploit the bettors' behaviour. The nominator of the Kelly fraction (Eq. (3.11)) is the so called "edge" or expected return of the bet. The edge considers both the correct prediction probability and the odds. When the odds are biased the edge decreases. Thorp (2008) argues that Kelly criterion may need millions of trials to dominate other strategies in case of having a low edge. On the other hand, Maclean *et al.* (2010) find Kelly criterion to be very risky in the short term and argue that despite its promising long-run growth properties, it may lead to poor return outcomes. In this Chapter's exercise, Kelly improves the betting performance in case of having enough trials, such as in OP where a bet is placed for each game. Similarly, when CF loses power and the forecasts deteriorate (as of the second year of OOS), Kelly can reduce the exposure to the risk and control the size of losses. On the other hand, for CF in the first year of OOS, where the model produces already significant profits, the Kelly criterion seems to have no effect on the results. This happens as the number of games of CF is applied is small and the model produces significant positive results that are not affected by the odds' biases.

This Chapter's empirical application has demonstrated the merits of CF. In a forecasting exercise where uncertainty and noise are high, CF has managed to generate forecasts that are accurate and profitable. These forecasts are generated from a transparent process that reveals also the factors that determine the target series. I note that although forecasting accurately the game result, number of goals and the result based on the Asian handicap do not seem a strenuous task with this Chapter's models, profitability is only obtained through CF. CF selects the forecasts of the underlying system that are strong enough for OOS estimation. That increases the level of accuracy to a level that is translated

to profitability. However, this level of accuracy naturally is being reduced in the second year of the OOS, to same cases below the level of profitability. In terms of the efficiency of the football betting market, I note that it is possible to generate consistent profits with CF. However, the practitioner needs to be adaptive and re-estimate their model at a frequent basis. At last, I note that in most cases, this Chapter's models are more accurate and profitable in the English Premiership. I assume that the English championship has less noise or in other words, it is easier to be forecasted<sup>26</sup>. At last, I demonstrate the benefits of the Kelly criterion and how it can be translated within the CF context. Its probabilistic nature seems highly beneficial in football betting in cases where the underlying model has a low power.

### 3.4 Conclusions

This study introduces the concept of CF and demonstrates its utility. CF generates a set of FRs in the IS and estimates their average firing strength. These rules are ranked and applied in the OOS. CF generates forecasting signals at points where strong rules are nearby and satisfy an endogenous and an exogenous threshold. The forecasting signal is a weighted average of the strongest rules. CF is useful in OR problems where uncertainty is high and poor forecasts are associated with substantial losses. It can offer transparency, protection against under and over-fitting while at the same time improves the forecasting accuracy of the underlying system.

In order to demonstrate its merits, a forecasting exercise is designed on the game result, Asian handicap and the number of goals of football games in the Premiership, La Liga and Seria A championships. CF is combined with RVM and generates forecasts in six consecutive seasons. These forecasts are evaluated in terms of statistical accuracy and betting profitability. In terms of the results, an active approach with higher frequency of retaining the forecasting models, improves the consistency of all models with the population dynamics. CF presents higher statistical accuracy and betting profitability than an RVM, RVM-ANFIS and

---

<sup>26</sup> It is well known amongst football fans, that in the English Premiership, teams' competition is higher. The number of teams that fight for the championship or to avoid relegation is higher than the one in La Liga and Seria A. Also, some of the biggest bookmakers originate from UK while football betting in the English Premiership is widely prevalent in Europe and Asia.

OP model. CF improves the accuracy of the underlying system (RVM) and RVM combined with ANFIS (the most common FIS). These results are translated into positive betting performance for the proposed procedure and negative for its benchmarks. The procedure outperforms not only the most popular fuzzy approach but also improves the predictability of the underlying system. This contrasts the common dilemma with developing transparent decision systems that sacrifice accuracy for interpretability (Martens, 2007). I also note that the Kelly ratio can further improve the profitability of CF and limit the losses of OP in the majority of cases.

CF can be a useful tool to practitioners and academics who deal with decision making problems that demand complex non-linear techniques. In areas such as medicine, finance and economics it can improve researchers' understanding of the underlying nature of the series. Where the public interest is concerned, in areas such as meteorology or football gambling, it can offer decision rules that are simple and interpretable. It is an attractive alternative to the numerous unconditional fuzzy inference approaches that dominate the literature.

## 4. Technical Analysis and the Discrete False Discovery Rate: Evidence from MSCI Indexes

### 4.1 Introduction

Technical analysis, commonly referred to as *Chartism*, is the type of investment analysis, which uses a class of graphical representations of financial asset's time series in order to explore trading opportunities. While technical trading is widely used by both investors and academics over the past century, there is a long and ongoing discussion on whether it genuinely has a predictive power and can generate sustainable profit in equities markets. Previous literature is split into studies highlighting the genuine profitability of technical analysis (see among others, Brock *et al.*, 1992; Hsu *et al.*, 2010) and those arguing against that (see among others, Sullivan *et al.*, 1999; Bajgrowicza and Scaillet, 2012).

Nonetheless, a comprehensive and up-to-date analysis of technical trading on equity indexes is still on demand from both academics and practitioners because the majority of previous studies tend to be narrowly focused on specific aspects of technical analysis on equity indexes. For example, a single market index, a restricted number and classes of technical trading rules, a “sterilized” exercise of technical analysis (e.g. no transaction costs involved) are totally different from what traders use in practice.

In the meantime, the use of a large universe of technical trading rules involves the appearance of the data snooping bias, which has been regularly investigated by the relevant literature. This issue has recently become widely known due to the enormous amount of data analysed by investors. The data snooping arises when a large pool of technical trading rules is exploited. The main concern is selection of certain rules whose performance are due to luck and so, not statistically significant.

Furthermore, even if technical analysis demonstrates significant predictability and excess profitability in specific markets and time periods, still there are several questions to be addressed. *What level of transaction cost can repel all market participants? Are there are some short-term anomalies in market efficiency allowing profitability? If yes, how long does this profitability persist*

and are the markets during these periods under stress or turmoil? What is the optimal IS to OOS ratio for achieving the best performance? These are the key research questions for this study answered by applying the proposed DFDR<sup>+/-</sup> method.

This Chapter conducts an extensive study of technical trading in the equities' markets. I study daily time-series of nine individual MSCI indexes and three general MSCI indexes replicating the performance of Developed, Emerging, and Frontier markets covering the period from 2006 to 2015. To analyse the technical trading, the Hsu *et al.* (2016) universe of 21,000 rules is considered. Their universe incorporates five main classes of technical trading indicators and oscillators.

Additionally, I propose a new method of controlling for data snooping bias while adjusting for the potential issues found in previous techniques. This Chapter's novel methodology is based on the False Discovery Rate (FDR) and specifically tries to expand the FDR<sup>+/-</sup> approach of Barras *et al.* (2010) in numerous ways. Their FDR<sup>+/-</sup> method is one of the most promising MHT<sup>27</sup> techniques. It can detect a sufficiently large number of statistically significant positive rules while allowing for a small number of false discoveries. When common resampling procedures (e.g. bootstrapping) are employed to compute each rule's corresponding *p*-values, a large set of homogeneous discrete *p*-values are realized, rather than uniformly distributed continuous ones (Storey, 2002; Storey *et al.*, 2004; Barras *et al.*, 2010; Brajgowicz and Scaillet, 2012). In addition, previous approaches of FDR<sup>+/-</sup> can lead to unnecessary conservativeness and consequently poor estimations of the proportion of rules. Such estimations rise the probability of Type II error where a significant rule is overlooked. The proposed DFDR<sup>+/-</sup> circumvents these issues by considering a large-scale homogeneous discrete *p*-values framework, while dynamically estimating the FDR hyperparameters. Hence, I provide a fully adaptive, computationally efficient approach to limit data snooping in the real world which can assist investors in analyzing their portfolios.

---

<sup>27</sup> The MHT deals with the problems where a large number of hypotheses are tested simultaneously. Testing the same rejection level for all hypotheses independently leads to a higher probability of making at least one Type I error compared to an individual test. The probability of having false inference increases exponentially as the number of studied hypotheses increase.



The proposed method is employed to perform several robustness checks and give answers to the research questions about the validity of the technical analysis. Specifically, an analysis of break-even transaction costs of the outperforming trading rules is exercised, while their OOS performance is investigated in a rolling-forward structure as fund managers do in practice. Moreover, the performance persistence of technical rules is analysed together with their performance during periods of turmoil by using stress measures from the Office of Financial Research (OFR). Finally, an innovative method for selecting the significant technical trading rules is introduced by cross-validating their performance between the full sample and IS and OOS subsamples.

In most problems in finance and economics researchers are dealing with a series of multiple competitive models or factors. In order to distinguish the genuine and significant ones, they most resort to MHT frameworks. The most common MHT approaches are the FWER, the FDR, and the False Discovery Proportion (FDP).

FWER is defined as the probability of having at least one Type I error. In other words, it measures the probability of having at least one false discovery. A testing method is said to control the FWER at a significance level  $\alpha$  if  $\text{FWER} \leq \alpha$ . Naturally, when a researcher performs a large number of hypothesis tests, it is highly likely to evidence at least one Type I error. There are several approaches to control the FWER. The most naïve method to control the FWER, is the Bonferroni correction. In this approach, the adjusted rejection zone is made by dividing the significance level  $\alpha$  over the number of tests. Then they run each test with a significance level  $\alpha/l$  (where  $l$  is the number of tests). The larger the number of tests, the smaller the common critical  $p$ -value. The Bonferroni correction is characterized by its simplicity but is criticized for loss of power and a high probability of Type II errors (Benjamini and Hochberg, 1995). A less strict approach to control the FWER, is the stepwise method of Holm (1979). For a set of  $l$  tests, the null hypothesis for the  $j$ -th  $p$ -value ( $p_j$ ) is rejected if  $p_j \leq \alpha/(l - j + 1)$ ,  $j = 1 \dots l$ . The criterion becomes less and less strict for large  $p$ -values and thus Holm's method rejects more hypotheses than the Bonferroni correction. However, both methods ignore the dependence structure of the individual  $p$ -values which makes them overly conservative.

White (2000) introduced the BRC to counter the problems of Bonferroni and Holm approaches. In his approach, the FWER is asymptotically controlled by estimating the sampling distribution of the largest test statistic and considering the dependence structure of the individual test statistics. BRC applies bootstrapping to get less conservative critical values than the previous approaches. The main limitation of BRC is that it only checks if the model or strategy that appears best within a set of candidates beats the benchmark. It also has low power when strong underperformers exist in the hypothesis testing pool and the  $p$ -values are still conservative (Romano *et al.*, 2008).

To address the problems of BRC, Romano and Wolf (RW) (2005) introduce the StepM test in an attempt to statistically validate as many outperforming strategies as possible. The RW test improves upon the BRC in a similar way the stepwise Holm method improves the single-step Bonferroni approach. The RW test initially identifies the most robust strategies through a step-down approach, until a false discovery is observed. The first step of the RW is the same as in the BRC test. In the next step, the remaining strategies are again evaluated over a new critical value (based on bootstrap) and these iterations continue until no further strategies are rejected. Given that RW is based on bootstrap estimates, it is safe to assume that it is less conservative (regardless the correlation structure of the  $p$ -values) and still asymptotically controls FWER (as BRC does). However, it still remains a strict approach since the procedure terminates once a false rejection is identified. To solve this issue Romano *et al.*, (2008) relax the strict FWER criterion by introducing the  $k$ -StepM method. The innovation of this method is that it allows for  $k$  number of false rejections before it stops compared to its predecessor. If the false selections are less than  $k$  the procedure continues in subsequent steps similar to RW. This makes the outcome less conservative, but the results are quite sensitive to the selection of  $k$ .

The FDR is based on the idea of allowing for a specific number of false negatives when a practitioner observes a quite large number of rejections, and by this way increasing the power of the test while relaxing the testing framework. Introduced by Benjamini and Hochberg (1995) as a more tolerant error metric, the FDR measures the proportion of false discoveries among true rejections of the null hypothesis. Specifically, they suggest that if  $F$  and  $R$  is the number of the total

Type I errors (false discoveries) and total null hypothesis (total discoveries) respectively, then the FDR is estimated as  $FDR = E(F/R)$ . Benjamini and Hochberg (1995) conclude that if all tested null hypotheses are true, then FDR is equivalent to FWER. However, if the number of true discoveries is lower than the total null hypotheses tested, then FDR is smaller than FWER. In addition, the FDR measures and controls the expected FDP, or in other words, it controls the FDR at level  $\zeta$  (i.e.,  $FDR = E(FDP) \leq \zeta$ )<sup>28</sup>. Over the years many studies have tried to develop further the FDR measure by improving the power and/or the adaptiveness of the tests. However, the common idea remains the same in all of them as identifying as many true rejections as possible without including too many false ones (Benjamini and Yekutiely, 2001; Storey, 2003; Storey and Tibshirani, 2003; Storey *et al.*, 2004, Liang and Nettleton, 2012; Liang, 2016). In financial applications, Barras *et al.* (2010) introduce for the first time an FDR approach similar to the one of Storey (2003) which focuses on measuring the proportion of false discoveries among mutual funds generating positive alphas, while trying to identify those displaying significant positive performance.

The relevant literature reflects the superiority of the FDR process (see among others, Harvey *et al.*, 2015; Bajgrowicz and Scaillet, 2012; Liang, 2016). The advantage of this method originates from the fact that by tolerating a certain (usually small) amount of Type I errors, the FDR improves the power of detecting more significant discoveries, compared to its stricter competitor: FWER. The FWER guards against a single erroneous selection and may lead to missed findings. The FDR approach uses lower critical values that allow a larger number of significant strategies to be selected. This is particularly important in finance and trading applications as investors prefer several alternative strategies, rather than constructing their whole strategy on a single trading tool. Additionally, the FDR test takes into account all outperforming rules in the population and it doesn't terminate when a single rule, even the best, yields a lucky performance. FDR is also able to identify a higher number of positive outperforming rules compared to the  $k$ -FWER (Psaradellis *et al.*, 2017).

---

<sup>28</sup>  $\zeta$  is user defined and should not be confused with  $\alpha$ .

The other pathway toward the MHT is FDP. This approach is based on controlling the probability of a user-specified proportion of false rejections for a single sample. This is comparable to the FDR approach that controls the expected proportion of false rejections across different samples. Controlling FDP can lead to more conservative estimates compared to the respective FDR ones (Genovese and Wasserman, 2006). Sun *et al.* (2015) suggest that in conditions of strong dependence FDP can be highly volatile. Also, Fan and Han (2017) report that true discoveries in FDP estimates can be relatively small for large datasets. Therefore, practically the FDR application is more suitable when analyzing large datasets and aiming to make confidence statements about the realized average FDP across the various datasets, as in this Chapter's case (Benjamini, 2010).

Based on the theoretical discussion above and characteristics of this Chapter's dataset, the FDR approach fits the scope of this Chapter's application better. This Chapter's dataset is a large universe of technical rules and I explore the dynamics of the MHT findings in different markets and periods.

The rest of this study is as follows. In Section 4.2, the theoretical grounds of the  $DFDR^{+/-}$  is discussed. Section 4.3 contains the details on the trading universe, the chosen stock markets and the performance metrics used to compare the pool of technical rules. Section 4.4 presents the characteristics for the set of true discoveries with an ex-post approach. Section 4.5 includes an ex-ante analysis with backtesting of the  $DFDR^{+/-}$  discoveries over the OOS. Section 4.6 provides the concluding marks. Finally, Appendix C includes a numerical comparison of the proposed  $DFDR^{+/-}$  with the RW through Monte Carlo simulations and studies the robustness of the methodology with a shorter IS period.

## 4.2 Methodology

### 4.2.1 Overview of the FDR Procedure

The FDR is defined as the proportion of false discoveries among true rejections of the null hypothesis. FDR is an expectation and thus its control does not require an additional specification on the probabilistic level (as in the FWER). Methods to control the FDR have been suggested by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001) and Storey (2002). The method of Benjamini and

Hochberg (1995) assumes that the  $p$ -values are mutually independent which is not plausible (discussed in Section 4.2.2). The Benjamini and Yekutieli (2001) approach assumes that  $p$ -values have a more arbitrary dependence structure, but it is less powerful. Storey (2002) improves its power with an approach based on the assumption that, for a two-tailed test, the true null  $p$ -values are uniformly distributed over the interval  $[0,1]$ , whereas the  $p$ -values of alternative models lie close to zero. His approach utilizes information from the centre of the test statistics' distribution – which is mainly dominated by non-outperforming rules – in order to correct luck in the tails. A key point towards this direction is the precise estimation of the proportion of rules satisfying the null hypothesis ( $\pi_0$ ) where  $\varphi_j = 0$ , in the entire population. A conservative estimator of the  $\pi_0$  parameter is given by

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda; j=1, \dots, l\}}{l(1-\lambda)} \quad (4.1)$$

where  $\lambda \in [0,1]$  is a tuning parameter indicating which specific level the null  $p$ -values exist. The required inputs for the FDR approach are mainly the (two-sided) corresponding  $p$ -values of the performance metrics ( $\varphi_j$ ) of each individual rule associated with the null hypothesis of no-abnormal performance ( $H_{0j}: \varphi_j = 0$ ) against the alternative of abnormal performance ( $H_{Aj}: \varphi_j > 0$  or  $\varphi_j < 0$ ). Furthermore, there is no need for a priori knowledge of the  $p$ -values distribution. The stationary bootstrap resampling technique of Politis and Romano (1994) is applied to obtain the individual  $p$ -values. It is applicable in cases where the time series are weakly dependent (which is the case in technical rules performance).

In this application, I am interested in identifying only the positive outperformers. In other words, the focus is on the case where  $\varphi_j > 0$ . For this reason, the  $FDR^{+/-}$  method of Barras *et al.* (2010) is incorporated to this Chapter's approach. In the context of technical trading rules performance, the  $FDR^+$  is described as the expected value of the proportion of erroneous selections,  $F^+$ , over the significant and positive rules,  $R^+$ , (i.e.,  $\frac{F^+}{R^+}$ ). The number of  $F^+$  represents the rules, whose  $p$ -values falsely reject the true null (i.e.,  $H_{0j}: \varphi_j = 0$ ) in favour of the alternative and exist among  $R^+$ . On the other hand,  $R^+$  portrays the number of rules rejecting the  $H_{0j}$ , in a two-tailed test, and their performance metric  $\varphi_j$

is positive. The estimate of  $FDR^+$  is given by  $\widehat{FDR}^+ = \hat{F}^+ / \hat{R}^+$  where  $\hat{F}^+$  and  $\hat{R}^+$  are the estimators of  $F^+$  and  $R^+$ , respectively. For example, an  $FDR^+$  of 100% conveys that no trading strategy genuinely outperforms the benchmark and any existing performance is purely out of luck. In general, the FDR produces a sensible trade-off between true positives and false selections, while having higher power compared to FWER. Due to its less conservative character, the FDR method is capable of identifying outperforming rules even if the performance of best rule is due to data snooping effect, contrary to previous methods.

Following Storey (2002), the frequency of false discoveries or the number of lucky rules,  $F^+$ , in the right tail of the distribution of performance metrics,  $\varphi_j$  at a given significance level  $\gamma$  is estimated as:

$$\hat{F}^+ = \pi_0 \times l \times \gamma/2 \quad (4.2)$$

where  $\pi_0$  is the proportion of rules satisfying the null hypothesis,  $\varphi_j = 0$ , in the entire population,  $l$  is the number of the entire population and  $\gamma/2$  is the probability of a positive non-genuine rule exhibiting luck due to symmetry conditions.

#### 4.2.2 Issues Regarding Existing FDR Methods

This section presents a discussion of possible issues arising in both the multiple hypothesis setup and the control of FDR based on previous methods. The procedure of Benjamini and Hochberg (1995) assumes that the multiple hypotheses tested are independent of each other. A considerable number of trading rules in this Chapter's trading universe are structurally similar to each other. For example, MAs, are highly correlated since a limited spectrum of parameters are considered to construct the universe. Several efforts have been made to address "weak dependence" conditions of the test statistics under which the FDR approach (Benjamini and Yekutieli, 2001; Storey, 2003; Storey and Tibshirani, 2003; Storey *et al.*, 2004; Farconemi, 2007; Wu, 2008). They argue that when the number of tests increases to infinity, the dependence effects diminishing to zero due to asymptotics. Likewise, in this Chapter's empirical investigation, the technical trading rules display dependence within specific classes (e.g. MAs),

while each class is independent to another as five different families of technical rules are considered. Therefore, I need to first check whether a weak dependence condition holds for this Chapter's dataset before further assessments.

Another critical problem about MHT is when for a large number of test statistics – usually in thousands – the number of observations is relatively small, just as in this Chapter's case. Specifically, I consider 21,195 technical trading rules over an IS horizon of two years (i.e., 520 observations). Bootstrapping procedures, as mentioned earlier, are commonly used in these cases to calculate the corresponding  $p$ -values and so to perform hypothesis testing. This is beneficial since they need a few distributional assumptions and they are robust to outliers. However, performing a resampling procedure on each trading rule, generates  $p$ -values which are discrete rather than continuous because of the finite number of bootstraps employed. This leads to the detection of large-scale homogeneous discrete  $p$ -values, sharing the same support points. Previous studies either controlling FDR or FWER, overcome this issue by assuming that the true null  $p$ -values are continuous and follow a uniform distribution as described above (Storey, 2002; Storey *et al.*, 2004; Romano and Wolf, 2005; Romano and Wolf, 2007; Barras *et al.*, 2010; Brajgowicz and Scaillet, 2012). Nevertheless, the true null discrete  $p$ -values tend to be stochastically larger than uniformly distributed and the direct application of existing methods on them can lead to misspecification (Pounds and Cheng, 2006).

Moreover, in a two-sided test and for continuous true null  $p$ -values uniformly distributed, holds  $Pr(p - value \leq \gamma) = \gamma$  for all  $\gamma \in [0,1]$ . Whereas for the discrete ones I observe only a certain number of support points (i.e.,  $V = \{\gamma_1, \dots, \gamma_v, \gamma_{v+1}\}$  with  $0 < \gamma_1 < \dots < \gamma_v < \gamma_{v+1} \equiv 1$ ) with potentially many ties. Using bootstrapping techniques to compute the associated  $p$ -values for each rule I end up with  $p$ -values satisfying a discrete condition with common support points. To illustrate further, every  $p$ -value is usually calculated by comparing the value of each performance metric with the value of its corresponding quantiles of bootstrapped metrics (Sullivan *et al.*, 1999). This means that, large values of observed test statistics provide evidence against the null and the corresponding  $p$ -values are given as  $p_j = \frac{1}{B} \sum_{i=1}^B (\varphi_{jb} \geq \varphi_j)$ .  $P$ -values computed this way are attached with support points of the form:  $V = \{\frac{1}{B}, \frac{2}{B}, \dots, \frac{B-1}{B}, 1\}$ , which also verify a discrete nature. Thus, an FDR

framework which takes into account large discrete  $p$ -values might help in improving further the existing methods.

Another issue appears in the calculation of  $\lambda$  parameter and so in the estimation process of  $\pi_0$ , which is the key estimator for controlling FDR. Generally, poor selection of  $\lambda$  can cause unnecessary conservativeness in  $\widehat{\pi}_0$  and of  $\widehat{FDR}$ . For example, not all values of  $\lambda \in [0,1)$  under a discrete setup generate ideal  $\widehat{\pi}_0$  estimates. For example, imagine a candidate set of  $\lambda$ ,  $\Lambda = \{\lambda_0, \lambda_1, \dots, \lambda_v\}$  with  $\lambda_0 \equiv 0$ . It can be shown that if one arbitrarily selects  $\lambda$  from  $\Lambda$  for some fixed element  $q \in \{0, \dots, v\}$  based on a support point, then  $\widehat{\pi}_0(\lambda)$  is a conservative estimator of  $\pi_0$  (Liang, 2016). On the other hand, if  $\lambda$  does not belong to this set but lies in between two support points, (e.g.,  $\lambda_i < \lambda < \lambda_{i+1}$ ), then choosing as  $\widehat{\pi}_0(\lambda) > \widehat{\pi}_0(\lambda_i)$ , can lead to an extra and unnecessary conservativeness in the estimation of the proportion of rules with no abnormal performance.

In terms of choosing  $\lambda$ , a small value can lead to estimators with large positive bias, while a large value of  $\lambda$  leaves only a small number of  $p$ -values on its right-hand-side to estimate  $\pi_0$ , yielding an increase on the variance of estimators. Thus, one should always achieve a good trade-off between the two when choosing  $\lambda$ . Previous literature follows a common approach to choose  $\lambda$  under a continuous set up; they visually examine the histogram of all  $p$ -values and set the  $\lambda$  parameter equal to the support point above which the number of  $p$ -value occurrences become fairly flat. The rationale lies on the assumption that bootstrapped  $p$ -values share equally spaced support points and each support contains a uniform number of true null  $p$ -values. A histogram approach might involve an extra bias towards on how the researcher conceives a specific level of a histogram's flatness. To address this issue, the next Section concentrates on selecting  $\lambda$  dynamically, as a fixed quantile of  $p$ -values, based on the data characteristics, while discarding any undesired conservativeness from the calculations.

Finally, MHT frameworks are computationally demanding most of the times since they involve bootstrapping procedures. Moreover, FDR approach requires to set the tuning parameter by taking into account the graphical representation of  $p$ -values, which considerably increases the computational time needed for



controlling FDR. This Chapter's proposed dynamic approach is computationally more efficient in terms of time and selecting the outperforming rules based on an algorithmic setup, which can also help practitioners making better decisions in portfolio construction and OOS estimation.

### 4.2.3 DFDR<sup>+/-</sup>

This Section presents this Chapter's new specification based on the FDR and homogeneous discrete  $p$ -values to identify outperforming trading rules while accounting for data snooping. This study is the first in the field of finance to propose an adaptive FDR approach employing a dynamic parameterization, while considering discrete  $p$ -values, as a tool for controlling data snooping.

This Chapter's approach concentrates on large-scale homogeneous discrete  $p$ -values. Following Kulinskaya and Lewin (2009), assume that by using the bootstrapping procedure (described in Section 4.2.2) I acquire discrete  $p$ -values, which satisfy a uniform condition while sharing the same discrete support  $V$ . Furthermore, I need to consider as  $N = \{n_1, \dots, n_{v+1}\}$  the number of occurrences of every component in  $V$ , i.e.,  $n_i = \#\{p_j = \lambda_i\}$  for  $i = 1, \dots, v + 1$  in order to express the empirical distribution of the computed  $p$ -values. Thus, the empirical distribution of homogeneous discrete  $p$ -values with common support points is thoroughly explained by  $(V, N)$ . Proceeding to the FDR approach calculation, the  $F^+$ ,  $R^+$  and  $FDR^+$  also represent step functions with possible change points at the support points. Then it is adequate only to acquire their values at the specific support points to control the FDR<sup>29</sup>. Given that, the distribution function of the null discrete  $p$ -values on every support point is almost identical to that of continuous  $p$ -values. This is the key step in extrapolating a method for discrete  $p$ -values from similar continuous  $p$ -values approaches.

This paragraph explains the novel approach to improving the FDR<sup>+/-</sup> methodology to accommodate discrete  $p$ -values, while dynamically selecting  $\lambda$  under a stopping time rule. I define this stopping time condition as the point in which holds  $E[\widehat{\pi}_0(\lambda_q)] \geq \pi_0$ , while  $q$  is the exact stopping time with respect to  $n_i$

---

<sup>29</sup> Suppose that  $\gamma$  is a time-running parameter from zero to one, then the continuous time processes  $F^+$ ,  $R^+$  and  $FDR^+$  relax to discrete stochastic process on the support points.

(for  $i = 0, \dots, v$ ), which includes the  $p$ -values up to  $\lambda_q$ , when  $q = i$ . I also determine the whole procedure up to the stopping time  $q$ , as  $\{0 \equiv n_0, \dots, n_q\}$ . Then I just set the effective  $\lambda$  equal to  $\lambda_q$ . I check every support point instead of checking every single  $p$ -value for the stopping condition. If  $q$  is an appropriate stopping time, it must also hold  $E[\widehat{FDR}(\lambda_q)] \geq FDR$ . The rationale of this approach is related to the idea of discovering the smallest support point, in which the number of appearances of  $p$ -values,  $n_i$ , to each right-hand side are almost equal. However, such a stopping time condition is very general as numerous stopping time rules can be designed fulfilling the above criteria since the true right-hand side counts are unobservable. The right-boundary procedure of Liang and Nettleton (2012) solves this issue by only considering the average of the remaining counts, been already known beforehand. The right-boundary specification guarantees conservative estimation for  $\pi_0$  and FDR. It relies on a grid of candidate points for  $\lambda$  in line with data characteristics and a stopping time condition, at least for a continuous space (Liang and Nettleton, 2012). I adopt the same approach and extend it to discrete  $p$ -values. It is worth noting that the right-boundary procedure performs effectively for both independent and weakly dependent  $p$ -values, as observed in this Chapter's case (see, Liang and Nettleton 2012; Liang, 2016). Liang (2016) provides evidence of computing an FDR estimator using the Discrete Right-Boundary (DRB) procedure, while certain limits exist. His results clearly satisfy a special case of the weak dependence condition of Storey et al. (2004).

The idea of DRB is to find the first  $\lambda$ , in which the values of  $\widehat{\pi}_0(\lambda)$  stop decreasing and satisfying in that way the stopping time condition. For this reason, I consider a candidate set for  $\lambda$ ,  $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ , in which I place its components in an ascending order,  $0 \equiv \lambda_0 < \lambda_1 < \dots < \lambda_n < \lambda_{n+1} \equiv 1$  (and  $\lambda \subseteq \Lambda$ ). Then I select the best  $\lambda$ , as the minimum  $\lambda_q$ , which fulfils that  $\widehat{\pi}_0(\lambda_i) \geq \widehat{\pi}_0(\lambda_{i-1})$ , (i.e.,  $q = \min\{1 \leq i \leq n + 1 : \widehat{\pi}_0(\lambda_i) \geq \widehat{\pi}_0(\lambda_{i-1})\}$ ). Specifically, I use the set  $\Lambda$  to separate the interval between zero and one,  $(0,1]$ , into  $n + 1$  bins with the  $i$ -th bin being  $(\lambda_{i-1}, \lambda_i]$  for  $i \in \{1, \dots, n + 1\}$  and  $w_i = \#\{p_j \in (\lambda_{i-1}, \lambda_i]\}$  being the number of  $p$ -values in the  $i$ -th bin. If the intervals between  $\lambda$ s are equal, then this approach actually chooses the right boundary of the first bin whose number of  $p$ -values is no larger than the average of the corresponding number to its right. In this way, I achieve the stopping condition when the downward trend of the number of  $p$ -values in each bin is neutralized, as I move forward, to a level which the random

variants of rest  $p$ -values are fairly equal. Finally, acquiring the optimal  $\lambda$  in this way, I can easily calculate a conservative estimator for  $\pi_0$  based on Storey's (2002) formula as have been already mentioned in previous sections.

The rest of the steps for the selection of outperforming rules remain similar with the Barras *et al.*, (2010) in the FDR specification. In terms of bootstrapping though, I generate 1,000 bootstraps of returns, and retain the same bootstrap draws of the time series sample period for each trading rule's returns. By this way, I bootstrap the cross-section of trading rules returns through time in order to preserve the cross-sectional dependencies (Kosowski *et al.*, 2006; Fama and French, 2010; Yan and Zheng, 2016). The application of stationary bootstrap also allows me to preserve the autocorrelations in returns structures. I then use the "point estimates" procedure of Storey *et al.* (2004) on generated  $p$ -values, under weak dependence to select the outperforming rules, while setting a target for false discoveries. I can also extrapolate the proportion of trading rules displaying nonzero performance as  $\pi_A = 1 - \pi_0$  in the entire universe of technical trading rules by using the FDR approach. This may be useful for an investor who wants to divide  $\pi_A$  into the proportions of positive,  $\pi_A^+$ , and negative,  $\pi_A^-$ , rules in the population. Appendix C.1 describes the precise steps of achieving this, the estimation of  $\lambda$  and so of  $\widehat{\pi}_0$ , as well as the computation of  $\pi_A^+$  and  $\pi_A^-$ . In this Chapter's Monte Carlo simulation presented – also in the Appendix C.1 – I provide evidence that this Chapter's  $\text{DFDR}^{+/-}$  procedure, achieves a good trade-off between the bias and variance in various weakly dependent settings.

#### 4.2.4 FDR Portfolio Construction

I construct portfolios of technical trading rules by setting the  $\widehat{\text{FDR}}^+$  equal to 10%, which achieves a good trade-off between the wrongly chosen rules and the truly outperforming ones according to the relevant literature. I find that results reveal a stability when  $\widehat{\text{FDR}}^+$  levels range from 5% to 30%. Hence, for the 10%- $\text{FDR}^+$  portfolio, 90% possess significant predictability while 10% of the rules selected do not have genuine predictive power among the outperforming rules. Moreover, I use the forecast-averaging technique and set equal weights to the signals pooled from the chosen rules to calculate the portfolio performance.

Each trading rule might generate a long, short or a neutral signal at a time-step. I invest an equal proportion of my wealth to the signals generated by each individual rule. Following previous studies (see among others, Brock *et al.*, 1992; Bajgrowicz and Scaillet, 2012), a trading position is opened when a long or short signal is produced and liquidated when the signal is either reversed or neutralized. Should a neutral position be raised, the fund is assumed to be invested in the risk-free asset or the saving account. The gross daily return is calculated by the change in the closing value of the underlying index. A one-way transaction cost is deducted from the gross return when a position is terminated. The excess return is then estimated to compare the profitability of the trading rules with the risk-free rate. The mathematical presentation of deriving the return time-series is explained in Section 4.3.3.

### 4.3 Dataset, Technical Trading Rules and Performance metrics

This Section presents the details of the environment where the DFDR<sup>+/-</sup> is applied. In Section 4.3.1 the dataset with the information regarding the studied markets is introduced. Section 4.3.2 describes the technical rules universe and Section 4.3.3 covers the performance measures used to compare the trading rules for different markets.

#### 4.3.1 Dataset

I study nine MSCI indexes that replicate the performance of United States (US), United Kingdom (UK), Japan, Brazil, China, Russia, Estonia, Jordan, and Morocco stock markets and the three categorical MSCI indexes that replicate the World (Developed), Emerging, and Frontier market indexes. The MSCI indexes are market capital-weighted indexes that reflect the holding returns of US investors in different markets. They are denoted in US dollars and are important references to institutional investors<sup>30</sup> (see among others, Hsu, 2010; Bena *et al.*, 2017). They include large and mid-cap segments of the benchmark markets and thus mitigate liquidity and tradability issues. The sample period for all time-series start from 1

---

<sup>30</sup> 99% of the top global investment managers are applying MSCI indexes (see, P&I AUM data and MSCI clients as of December 2017).

January 2004 and end on 31 December 2016. The summary statistics of the log returns of the twelve series and of the risk-free rates series are presented in Table 4.1.

[Table 4.1]

All indexes are leptokurtic while the risk-free rate series is behaving very close to the normal curve. UK, Brazil and Russia have very high kurtosis. All time-series except for the Frontier index and the risk-free rate exhibit negative skewness (the UK has the least). The positive autocorrelation coefficient is seen for all times series except for Japan and US; however, the reported coefficient is not statistically significant for Japan.

### 4.3.2 Technical Rules

Technical trading aims to recommend long or short positions for the next period based on historical quotes for open, high, low and close prices along with other characteristics such as previous trends, momentums and directional movements. In this study, 21,195 rules are utilized following the work of Hsu *et al.* (2016). This universe of trading rules includes FIRs, Relative Strength Indicators<sup>31</sup> (RSIs), MAs, S&Rs and CBs. These technical indicators are common in practice and are available in trading websites, numerous research papers and textbooks in Finance.

Appendix A.1 presents short descriptions of FIR, MA, S&R and CB rules. For the exact characteristics of the studied technical trading universe, the reader is referred to Hsu *et al.* (2016). In total, 21,195 technical rules are generated for each of the twelve MSCI indexes under study.

### 4.3.3 Excess Returns, Transaction Costs and Performance metrics

In this Section, I define the daily excess return for every index examined as well as the performance metrics employed in accounting for transaction costs. Firstly,

---

<sup>31</sup> RSIs are momentum oscillators that measure the speed and change of price movements. Momentum is measured as the ratio of higher closes to lower closes: stocks with more or stronger positive changes have a higher RSI than stocks which have had more or stronger negative changes. RSI is considered overbought when above 70 and oversold when below 30.

I calculate the daily gross return from buying and holding the index during the prediction period as:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (4.3)$$

where,  $\ln\left(\frac{P_t}{P_{t-1}}\right)$  is the daily gross return from buying the pair and holding it for one day,  $P_t$  is the spot price on day  $t$  and  $P_{t-1}$  is the spot price on the previous day. Each calendar year is assumed to have 260<sup>32</sup> trading days on average.

Secondly, I need to consider the impact of transaction costs into the technical trading simulation. For that reason, I treat transaction costs “*endogenous*” to the trading process. For instance, I deduct one-way transaction costs every time a long or short position is closed according to the next period’s index value prediction. I estimate the one-way transaction costs taken at time  $t$  for trading rule  $j$  as:

$$TC_{j,t} = I_{j,t} \times tc \times P_t \quad (4.4)$$

where,  $I_{j,t}$  is the indicator set to 1 when a position is closed for the studied trading rule and the transaction cost  $TC_{j,t}$  is deducted (0 otherwise) at time  $t$  and  $tc$  represents the level of transaction costs used.

The transaction cost can negatively affect the performance of the portfolios (Cesari and Cremonini, 2003). Industry-based factsheets along with the academic literature recommend a transaction cost of 25-75 basis points (bp) for trading MSCI indexes (Cesari and Cremonini, 2003; Investment Technology Group, 2013; Eurex, 2018). MSCI (2013) suggests transaction costs up to 50 bp for their indexes. The same survey shows that higher transactions cost (75 bp) leads to making the indexes out of money. In this study, I consider a one-way proportional transaction cost of 25 bp for advanced markets (US, UK, Japan and Developed) and 50 bp for

---

<sup>32</sup> In finance literature each calendar year is expected to have 252 trading days. However, since markets operate on different schedules around the world the number of trading days differ in this case.

the other markets. These costs correspond to fees, bid-ask spread and slippage. These costs are realistic for large institutional investors.

In terms of performance metrics, I provide the annualized *mean excess return* and *Sharpe ratio* criteria. In this way, I consider an absolute criterion based on each technical trading rule's returns – the mean excess return – and a relative performance criterion reporting the ratio of the mean excess return to the total risk of the investment in terms of excess returns' standard deviation – the Sharpe ratio. Consider the trading signal  $s_{j,t-1}$  triggered from a trading rule  $j, 1 \leq j \leq l$  (where  $l = 21,195$ ) at the end of each prediction period  $t - 1$  ( $\tau \leq t \leq T$ ), where  $s_{j,t-1} = 1, 0, \text{ or } -1$  represents a long, neutral or short position respectively taken for time  $t$ . The mean excess return criterion  $\bar{f}_{j,t}$  for the trading rule  $j$  is given by:

$$\bar{f}_{j,t} = \frac{1}{N} \sum_{t=\tau}^T [s_{j,t-1} r'_t - TC_{j,t} - \ln(1 + r_{f,t})], \quad j = 1, \dots, l \quad (4.5)$$

where  $N = T - \tau + 1$  is the number of days examined and the term  $[.]$  is the excess return net for the risk-free rate  $r_f$  at time  $t$ . The  $\tau$  is the activation period since some of the technical trading rules use lagged values of indexes up to one year (260 days). For the risk-free rate, I use the effective federal funds rate reported by Federal Reserve in the US. Since the quotes for the risk-free rate are reported on an annual basis, I transform the rates to the daily basis as,  $r_{f,t} = (1 + S_t)^{\frac{1}{260}} - 1$ , where  $r_{f,t}$  is the estimated daily rate and  $S_t$  is the quoted federal funds rate.

Finally, the Sharpe ratio metric expression  $SR_j$  for trading rule  $j$  at time  $t$  is defined by:

$$SR_{j,t} = \frac{\bar{f}_j}{\widehat{\sigma}_j}, \quad j = 1, \dots, l, \quad (4.6)$$

where  $\bar{f}_{j,t}$  and  $\widehat{\sigma}_{j,t}$  are the mean excess return and the estimated standard deviation of the mean excess return respectively. An important feature of the Sharpe ratio metric is its direct link with the actual test statistic of the empirical

distribution of a rule's returns (Harvey and Liu, 2015)<sup>33</sup>. This theoretical connection together with popularity of the Sharpe ratio in trading industry makes the Sharpe ratio the most appropriate criterion for this Chapter's proposed MHT framework<sup>34</sup>.

Through Eq.s (4.5 and 4.6) each rule's performance metric ( $\varphi_j$ ) is calculated and tested for significant positive difference compared to a benchmark. Following Sullivan *et al* (1999) and Bajgrowicz and Scaillet (2012), this Chapter's benchmark is the risk-free rate that corresponds to abstaining from the market when no profitable opportunity is expected. Alternatively, the benchmark can be defined as the buy and hold strategy on the MSCI World index or further to a combination of bonds and stock indexes<sup>35</sup>.

## 4.4 IS Performance

This Section provides an ex-post analysis for the technical rules. In Section 4.4.1 the number of strategies identified as genuine profitable in the IS and their relevant profitability is presented. Section 4.4.2 reports a break-even transaction cost analysis. The corresponding results based on one-year IS are presented in Appendix C.2.

---

<sup>33</sup> The test statistic of a given sample of historical returns ( $r_1, r_2, \dots, r_t$ ), testing the null hypothesis that the average excess return is zero, is usually defined as  $t = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}}$ , while the corresponding Sharpe ratio is given by the formula  $SR = \frac{\hat{\mu}}{\hat{\sigma}}$ . The statistic  $t$  was used in Chapter 2 to compare the trading universe.

<sup>34</sup> Financial literature has a wealth of alternative performance assessment measures e.g. the Sortino ratio (Sortino and Price, 1994) and Manipulation-Proof Performance Measure (MPPM) (Goetzmann *et al.*, 2007). Although the results vary with the alternative choice of the test statistic, similar trends are observed with the Sortino ratio and the annualized excess return as performance metrics.

<sup>35</sup> The choice of a relevant benchmark is central to the hypothesis testing and the set of discoveries. Although different possible specifications could be considered, the scope of this study is limited to proposing the new MHT procedure in the most common and verified setting. Exploring different choices of the statistic and the benchmark remains for other studies such as the application in Chapter 5.



### 4.4.1 Identification and IS Profitability

Table 4.2 presents the percentage and standard deviations of the survivor rules identified by the 10%-DFDR<sup>+</sup> approach. The IS periods are set to two years and the portfolios are re-adjusted on a monthly basis.

#### [Table 4.2]

I note that the percentage of identified rules varies throughout the years and the markets. The higher number of identified rules are in the UK, Russia and the Frontier markets indexes. There is no obvious trend in the percentage of identified rules in terms of years. The peaks are in 2006, 2009, and 2010 while the lowest is in 2012. It is interesting to note that this Chapter's procedure identifies profitable rules for all indexes and all years. This means that the null hypothesis of no positive Sharpe ratio is rejected in all cases. The relevant trading performance (Eq. (4.5)) of the portfolios generated in the IS is presented in Table 4.3.

#### [Table 4.3]

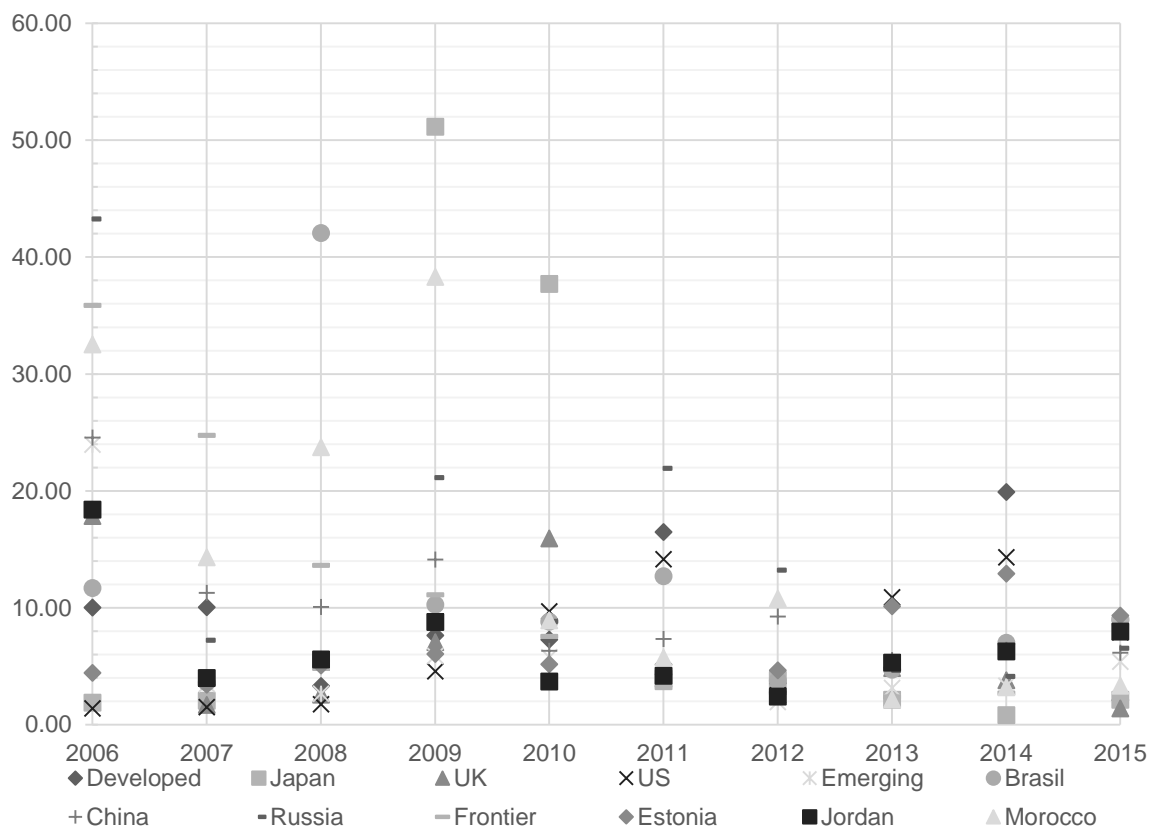
There is significant profitability after transaction costs for all indexes. Emerging markets present an increased profitability compared to their counterparts in terms of annualized return. There is no obvious trend on the profitability of technical analysis. There is a peak for the years 2009 and 2010 and consistent stable Sharpe ratios for the following years. There is also no connection between the percentage of identified rules and the trading performance of the generated portfolios.

### 4.4.2 Break-even Transaction Costs

In this Section, I perform a break-even analysis of technical trading rules' excess profitability over the IS period. Following relevant studies in the field (see, Bessembinder and Chan 1998; Bajgrowicz and Scaillet 2012; Hsu *et al.*, 2016) I adopt as a break-even cost the size of one-way transaction costs, which make the excess profitability (i.e. mean return) of the best-performing technical trading rule to diminish to zero. The more the break-even costs surpass the actual ones the more robust a rule's excess profitability is assessed.

Figure 4.1 displays the size of average break-even transaction costs (in percentage) for the best-performing technical trading rule under the Sharpe ratio metric and for each index separately. I select the best-performing rule for every month based on the previous two years period that acts as the IS. Then the procedure is repeated for all 12 months per year, while rebalancing is performed. The average break-even transaction cost per year is estimated by dividing the sum of the best rule's monthly break-even transaction costs by twelve. The same procedure for the overall 10-year period is applied.

**Figure 4.1. Break-even Cost for the Top Performing Survivor of the DFDR<sup>+</sup> Procedure (IS 2 Years)**



**Note:** The values are in percentages and calculated as the transaction cost that sets the excess return to zero over the period under study. The IS period is set in two years, while the same results for IS of one year are available in Figure C.1. The values are calculated by repeating the procedure at the start of each month and averaging over 12 months.

The major trend reveals by the figure is that frontier markets achieve the highest break-even transaction costs, with second and third best the emerging and developed markets respectively, at least for the first four years. In particular, Morocco (i.e., 33-38%), Russia (i.e., 21-43%) and Brazil (i.e., 12-42%) dominate in terms of excess profitability robustness over that period, while on the other hand an advanced market, namely Japan (i.e., 52%) reports the highest break-even costs over 2009 and 2011. For the rest over the years, there is a decay of break-

even transaction costs, except for 2014, in which case the developed markets recover compared to the rest. For instance, the break-even transaction costs of the corresponding developed markets' index as well as the US span from 14% to 20%. Undeniably, there are still some emerging (i.e. Russia and Brazil) and frontier (i.e. Estonia and Morocco) markets, which score similar or even higher break-even costs.

In general, I observe a downward trend of break-even costs and so excess profitability over the years, reaching their lowest levels in recent years and especially in 2015. However, this trend is not stable all the time. Figures 4.1 and C.4 (see Appendix C) reports a cyclicity in technical trading rules profitability, which is more consistent with the AMH. Specifically, most of the countries exhibit high break-even costs on 2006 facing a small decay in 2007, while they return to higher levels from 2008 to 2010, in which years they also reach a characteristic peak. During the following years, there is a considerable decay on their size, with a slight recovery only in 2014, which doesn't clearly remain on 2015. So, except the years in which break-even costs report their highest values (i.e. 2008-2010), there is a relatively consistent performance of technical trading rules, especially in recent periods.

## **4.5 OOS Performance**

This Section provides a comprehensive analysis for the portfolios constructed on the surviving rules of the DFDR<sup>+</sup> procedure with an ex-ante approach. Section 4.5.1 presents the excess profitability of the DFDR<sup>+</sup> portfolios over the OOS. Section 4.5.2 studies the performance persistence by measuring the number of periods a DFDR<sup>+</sup> portfolio can generate a return above the risk-free rate. In Section 4.5.3, a novel cross-validation practice is provided that considers both IS and OOS and preserves the order of the underlying time-series. Finally, Section 4.5.4 focuses on interpreting the excess returns of the portfolio over the different period by considering the level of financial stress in markets.

### **4.5.1 Excess Return**

Following previous studies in the field, I employ an OOS experiment based on the significant technical trading rules found by the DFDR<sup>+</sup> procedure. I create

portfolios of outperforming trading rules for each index based on their IS performance and evaluate them in OOS. I consider three OOS horizons – one, three and six months – following a two-year IS period<sup>36</sup>. For instance, in the case of one-month OOS, I select the significantly positive rules based on their previous two year's performance tested by the 10%-DFDR<sup>+</sup> approach as my portfolio selection tool (see Section 4.2.4). Then the portfolio's performance is evaluated in the following one-month period. I rebalance the DFDR<sup>+</sup> portfolio every month in a rolling-forward structure within a year and I repeat the same procedure for all the years in this Chapter's study period (i.e. 2006-2015). By this way, I dynamically build and evaluate the portfolios like a profit-seeking investor. I utilize the OOS periods of three and six months in a similar manner.

Table 4.4 reports the average excess annualized mean returns (Eq. (4.5)) followed by the Sharpe ratios (Eq. (4.6)) in parenthesis. For every index, a two-year IS and a one-month OOS are considered. The values in Table 4.4 are calculated as the corresponding OOS averages of twelve DFDR<sup>+</sup> portfolios built for every index after rolling-forward the IS by one month during one-year time span.

#### [Table 4.4]

OOS evidence shows that technical trading portfolios outperform in almost all markets during the earlier years and especially during 2006. In terms of developed markets, the UK and the US MSCI indexes both demonstrate positive mean returns and Sharpe ratios. Regarding the emerging markets, excess profitability is even more profound, with all three markets (i.e. Russia, China and Brazil) providing outstanding mean returns and Sharpe ratios (e.g., 59.59% and 2.62 respectively for China). A similar performance is observed for the case of Morocco. However, this trend does not seem to persist for 2007 where only a limited number of the emerging (China and Brazil) and the frontier markets (Morocco) retain their positive performance. Focusing now on the next years and especially in 2008 and 2009, almost all markets present extraordinary performance, which is more solid for frontier markets compared to the others.

---

<sup>36</sup> I also set the IS period covering one year, however the corresponding portfolios of significant rules perform slightly worse than those constructed with a two-year IS period. The relevant tables and discussions are provided in Appendix C.2.

The years 2008 and 2009, corresponding to the global financial crisis period, during which most of the markets faced extreme downward trends and experiences severe losses. This environment seems beneficial for this Chapter's technical trading portfolio as it mainly consists of momentum rules that can use such big trends. This also explains the negative performance for most cases in 2010, which was a turning point year for most of the markets. Additionally, the under-performance of the DFDR<sup>+</sup> portfolios continues for the rest of the years, signifying decay on technical trading rules performance. There are only a few exceptions (i.e. UK, US, China, Estonia and Morocco) on specific periods in which technical trading can still achieve some returns in over the recent periods but even these do not seem to be retained for more than one year. Considering the categorical MSCI indexes (Developed, Emerging and Frontier) the DFDR<sup>+</sup> portfolios' performances are not generally consistent with the corresponding performance of portfolios on their constituent indexes (e.g. Developed compared to the US). Only Frontier markets index provides a more solid positive performance over the study period, generating positive mean returns and Sharpe ratios even in the latest years.

Table 4.5 displays the average excess annualized mean returns and Sharpe ratios (in parenthesis) after transaction costs for every index using the same two-year period as IS but a three-month as OOS period. All values are computed as the OOS averages of the twelve portfolios per annum for every index in rolling-forward structure.

#### [Table 4.5]

The main finding in Table 4.5 is the similar pattern with those of using one month as the OOS period, with most of the developed and all of the emerging markets and Morocco from frontier markets to demonstrate healthy performance metrics in 2006. A performance, which is only preserved by some of the emerging markets (i.e. China and Brazil) and Morocco in 2007. During the global financial crisis period (i.e. 2008-2009) the picture is the same as in Table 4.4, with technical trading rules outperforming, but this time mainly during 2008. Consequently, four out of the nine country-specific markets switch to negative returns in 2009. This might be a result of a longer-term OOS period examined and failure to tuning for new trends. Once again, the performance of technical trading rules diminishes

over the most recent periods, however, almost the same countries as before (i.e. US, UK, China and Estonia) demonstrate excess profitability on specific periods. This may indicate specific patterns in market trends during these years, exploited by technical trading rules. In terms of the categorical MSCI indexes, the most promising index remains the Frontier index. However, the robust excess profitability for the Frontier index over most of the years is no longer stable as in Table 4.4. In almost half of the years, the Frontier index yields negative returns.

Finally, Table 4.6 presents the OOS performance of DFDR<sup>+</sup> portfolios by applying a six-month post-sample period in a rolling-forward structure.

#### [Table 4.6]

As in the case of one and three months OOS periods, there is a considerable evidence of excess profitability of technical trading rules across most of the markets examined during the earlier years, while it reaches its highest level in 2008. A considerable decay is observed over the more recent years, but some level of excess profitability remains for specific indexes and periods. Interestingly, I observe that the profitable indexes have been reduced almost to half, especially during profitable years such as 2008, compared to Tables 4.4 and 4.5, in which a smaller post sample period is applied. However, technical trading is still able to identify and predict quite well some indexes based on the evidence of all three OOS periods considered. For example, US and UK indexes in 2013 and 2014 from developed markets, China index in 2015 from emerging markets and Estonia in 2014 from frontier markets yield consistently positive performance metrics. This signifies the presence of some patterns, which can be captured by technical analysis. Regarding the categorical indexes, two out of three, namely the Emerging and Frontier indexes generate quite healthy mean returns and Sharpe ratios in some periods, which span even to recent years.

Analysing Tables 4.4 to 4.6, the performance of technical trading rules in emerging and frontier markets is stronger compared to the developed markets at least in profitable periods (i.e. 2006 and 2008) during which the former ones achieve higher mean returns and Sharpe ratios. This is also validated when the OOS period is changed. However, none of the markets seems to be consistent in terms of performance over the entire study period.

### 4.5.2 Performance Persistence

The results in Section 4.5.1 recommend that technical rules and this Chapter's approach can benefit from short-term market inefficiencies. However, market efficiency should diminish any profitability sooner or later. In this Section, I check how fast this profitability decays and whether there are anomalies between the different markets and periods. This element is overlooked in the related literature where the empirical evaluation is static and limited to specific periods. In real-world trading environments, practitioners are adaptive and rebalance their portfolios on a frequent basis.

Table 4.7 presents the persistence of this Chapter's generated portfolios for the two-year IS and one-month OOS case. I measure persistence as the number of consecutive months for which my portfolios have a trading performance above the relevant risk free-rate.

#### [Table 4.7]

I note that in most cases, traders need to rebalance the portfolios on more than a monthly basis. These results are expected considering the level of efficiency of financial markets and the mediocre trading performance of the trading rules in the previous section. However, it should be noted that there is a handful of cases where the measured robustness is higher than unit. In other words, there are cases where market efficiency was weak enough to allow profitability for static portfolios. Persistence is higher for the years 2007 and 2008 that can be attributed to the turbulence due to the global financial crisis and its side-effects to market efficiency. Interestingly, the market with the higher persistence is the US, the largest and more liquid index under study. In Table 4.8, I repeat the same exercise for the IS 2 years and OOS 3 months case.

#### [Table 4.8]

The persistence is decreased for most cases and years. I note that in 2008 the average persistence of my portfolios is more than one. Else, my portfolios retain their profitability after the first three OOS months. I note that the US retains a persistence larger than one on average. This is surprising as someone

might expect frontier and emerging markets to have stronger persistence than developed markets. Table 4.9 presents the same exercise for IS 2 years and OOS 6 months.

### [Table 4.9]

From Table 4.9, the further persistence decreased is observed. There are years (especially for frontier markets) where persistence is zero. It means that none of the generated portfolios has a trading performance higher the relevant risk-free rate for the first six months of the OOS. Similar to the previous two cases, I note a peak for the year 2008.

The results in this Section highlight the importance of holding periods in trading, a fact overlooked in the related literature. I note that there are a few cases where the portfolios might have a negative profitability in the first month of the OOS but some of them bounce back in the following periods (see, for example, persistence for the year 2008 in Tables 4.7 to 4.9). In these cases, adaptiveness seems not to always lead to increase profits while patience is often rewarded. However, the majority of the results highlight the importance of rebalancing the portfolios. Emerging and frontier markets do not seem to offer a safe haven to static portfolios. The observed trends allow me to remark that persistence is stronger at the peak of the global financial crisis of 2008. These results lead to further explore whether financial stress levels affect the profitability of technical analysis (see, Section 4.5.4). This exercise also highlights the importance of choosing the IS and OOS length. This choice can be seen as a trade-off between Type I and Type II errors on the modelling part (Harvey and Liu, 2015). On the empirical side, it can be seen as how adaptive a trader should be and how persistent are the technical rules.

### 4.5.3 OOS Cross-validation

Backtesting to measure the performance of a trading strategy over the unseen market conditions is a common approach for both academia and industry. However, it can lead to false inference sometimes. For instance, an OOS simulation doesn't always represent a true futuristic replication since the practitioner is already informed about what happened in the economy. In other



words, the outcome is always realized beforehand. Additionally, the data splitting in IS and OOS horizons plays an important role in the success of OOS experiment and the performance of the trading strategies over specific periods (see Sections 4.5.1 and 4.5.2). Furthermore, it is very likely to overlook true alternative (i.e. false negatives) when only a part of the full sample data is reserved for performing an IS-OOS backtesting.

In this Section, the robustness of the OOS performance is revisited in an innovative way based on a cross-validation experiment between the findings of IS and OOS. I explore a method proposed by Harvey and Liu (2015), which involves a combination of the full sample and the IS-OOS evidence in order to search for the intersection of survivors. If the IS period is not too short (i.e. two years) compared to the OOS corresponding periods (i.e., one, three and six months, respectively) I retain the IS-OOS test results computed previously. Those involve the genuine technical trading rules survived in OOS after employing the DFDR<sup>+</sup> approach at 10% target of false rejections and while considering the Sharpe ratio metric as this Chapter's test statistic. In this case, I also employ the DFDR<sup>+</sup> method to select the significantly positive rules for a full sample horizon with a more lenient target rate (i.e., 20%). I consider three full sample periods corresponding to each of the three different OOS periods examined plus the IS period, while I utilize the aggregate dataset each time (e.g. two years and one-month full sample in the first case). Then, I merge the findings observed by the IS-OOS and the full sample simulations in order to identify the potential intersection of significant rules provided by the two approaches. Finally, I evaluate the performance of the rules belonging in the intersection. At my best knowledge, this is the first time of such an approach is deployed.

Tables 4.10 to 4.12 report the results of the cross-validation test as described above. The tables report the average OOS annualized mean excess returns of twelve cross-validated portfolios. The portfolios are constructed similar to the previous sections (4.5.1 and 4.5.2) while accounting for transaction costs together with the average percentage of cross-validated rules out of the total number of surviving rules – presented in Table 4.2 – (in parenthesis) are reported.

[Table 4.10]

Table 4.10 exhibits the results of my cross-validation exercise for the one-month OOS case. Profitability is very high, especially during the global financial crisis (i.e., 2008-2009). For the rest of the periods, emerging markets demonstrate the highest performance and report very healthy returns with only a slight decay during the most recent years. The frontier markets exhibit a similar pattern in profitability. For developed markets, the performance of technical trading rules seems promising over most periods, while the returns yielded are quite lower compared to the ones of developed and frontier markets. Positive profitability was naturally expected, as cross-validated rules are a subset of the profitable OOS rules. The percentages of cross-validated rules vary considerably. For the developed indexes, these numbers span from 0.01% to 17.41% (both for the US), while for emerging indexes range from 0.01% (China and Brazil) to 18.07% (China) across all periods. Moreover, the relevant percentages of cross-validated rules for frontier indexes spread from 0.01% (i.e., Morocco, Jordan) to 18.14% (i.e., Morocco). On average, the percentage across all markets and periods is 2.47% or 30 rules. This number demonstrates the value of technical analysis in trading. Table 4.11 has the results of the exercise for the 3 months OOS case.

**[Table 4.11]**

I observe the same characteristics in technical trading rules performance; however, the magnitude of generated returns are smaller compared to the ones obtained in the one-month OOS period. The percentage of the cross-validated rules is on the same levels as the previous case. Table 4.12 presents the 6 months OOS case.

**[Table 4.12]**

The evidence provided by performing the cross-validation experiment in a six-month OOS horizon reveals even lower returns compared to the relevant ones using one and three months across all indexes and overall post-sample periods. This finding is consistent with the one in Section 4.5.1 describing the OOS results profitability since the significant rules are exposed to longer post-sample periods uncertainty. In terms of profitability patterns, all market indexes are able to achieve high returns during the earlier years (i.e., 2006-2009), with emerging indexes being more profitable, while frontier and developed markets come second

and third respectively. However, developed market indexes' returns seem to diminish over the recent periods and especially after 2008 and 2009 in which years they achieve their highest returns. Regarding the emerging market indexes, technical trading returns remain more robust and yield high returns even in recent years (i.e., 2014, 2015), while no specific pattern is observable at least for the Russian and Brazil indexes. On the contrary, the considerable decay in profitability is more profound in frontier markets indexes from 2009 to 2010 relative to the developed indexes. These findings are generally the same for the corresponding categorical MSCI market indexes. In addition, the average percentage of cross-validated rules has slightly increased compared to the relevant amounts of the three-month OOS period. For instance, these percentages reach 25.83% (US) for developed indexes, while they have moved up to 15.23% (China) and 20.98% (Morocco) for emerging and frontier indexes respectively.

This Section finds that a limited number of genuine profitable technical rules exist. The DFDR<sup>+</sup> approach can identify subsets of these rules as demonstrated in Section 4.4.1. The percentages of cross-validated rules might seem small, but it should be noted that the scope of this Section is to check whether truly profitable rules at both the IS and OOS do exist. This exercise allows me to suggest with further confidence that technical analysis has value in trading. The results can also be seen as the performance of an "oracle" trader that applies technical rules in studied dataset.

#### 4.5.4 Financial Stress

The previous sections note a peak on the technical rules' performance for the years 2008 and 2009 that correspond to the recent global financial crisis. The performance of the portfolios deteriorates in the following years, but there are still cases where they present excess profitability after transaction costs even in developed markets. This Chapter's results contradict the previous recent literature that finds no recent excess profitability after transaction costs for technical analysis (see among others, Hsu *et al.*, 2010; Bajgrowicz and Scaillet, 2012; Taylor, 2014)<sup>37</sup>. The authors argue that the popularity of Exchange-Traded

---

<sup>37</sup> As discussed before, none of the previous literature presents an extensive empirical application as the one presented in the previous sections. Previous authors also applied more conservative and time consuming MHT frameworks and limited their empirical applications to specific years or

Funds (ETFs), algorithmic trading, market liquidity, derivatives or the effect of other macroeconomic factors have eliminated excess profitability in recent years. Although the effect of these factors cannot be discarded, the results suggest exploring the effect of financial stress on the portfolios.

The literature is rich in financial stress indexes. The difference is based on the components that are used to construct them, their frequency and the market that they are applied to. In this Chapter's study, I apply the OFR stress daily indexes. They are constructed for 33 market financial variables and cover the US, other developed economies and emerging markets<sup>38</sup>. I match the findings for the US, Developed and Emerging indexes from Section 4.5.1 with the stress levels as reported by the related US, other developed<sup>39</sup> and Emerging markets' OFR stress indexes. The stress levels are averaged based on the daily OFR indexes for each month. A month is then classified as either high or low stress. Then for each year, the trading performance over the high- and low-stress is measured separately. Table 4.13 presents the performance of my portfolios under high and low financial market stress.

### [Table 4.13]

The table offers two main findings that highlight the complexity of the financial markets. Firstly, there is no solid verdict about the stress. Speculators might consider benefiting from disruptions in the normal financial market activities over highly stressed periods. This can lead to a higher reward but might also cause unexpected losses. For example, consider the case of 2010 – the year in which most markets present negative trading performance in Section 4.5.1. A

---

periods. For example, let me consider the case of the MSCI US index, the 2 years IS and 1-month OOS (see, Table 4.4). If the application was limited only to year 2014 or 2015, the interpretation would be unrealistic. The flexibility and adaptiveness of DFDR<sup>+/−</sup> along with the recent developments in computational power allowed me to conduct an empirical analysis that unveils previously unknown patterns.

<sup>38</sup> Other indexes (such as the St. Louis Fed Financial Stress Index and the Kansas City Financial Stress Index) focus on a specific index while others stop before my sample (such as the IMF stress index). My criteria to select the index are to cover as many markets as possible from the ones under study and to apply the all indexes to been constructed with the same methodology.

<sup>39</sup> I note that the MSCI and the OFR indexes don't match perfectly. For example, the "other advanced" OFR index does not include US. However, among the more cited financial stress indices, OFR is the closest to my study.

trading practice over the high-stress period exclusively in the US market could almost hedge the loss while the same practice for the Emerging index could lead to a 20% loss. Therefore, the decision to invest in such conditions depends on the risk preference of the investors. Secondly, comparing the average performance of different trading practices (high, low and the original) provides mixed results. Financial markets operate continuously and digest the inbound flow of information on a regular daily basis. Fundamental developments are not priced in financial markets as expected by central banks (Kontonikas *et al.*, 2013). Stress indexes like the OFR index try to cover the fundamental factors. Segregating the markets into high-stress and low-stress submarkets based on such measures does not necessarily provide a superior market insight.

## 4.6 Conclusion

This study uses a novel MHT approach to measure the profitability of technical analysis. The  $DFDR^{+/-}$  is a simple and efficient tool for comparing a large-scale set of inputs with structural dependence and discrete  $p$ -values. Specifically, the FDR approach is used to assess a technical trading universe from five main families of oscillators and indicators. The tested null hypothesis is that under an unbiased evaluation, different variants of the technical rules are practically the same and provide no significant profit.

A thorough experiment is designed to quantify the merits of the  $DFDR^{+/-}$  method and test the validity of technical rules. Monte Carlo simulation is selected to compare the proposed method with an FWER benchmark under deterministic conditions. After validation by random data, the  $DFDR^{+}$  tests the null hypothesis for the trading universe. Then, significantly positive rules based on the  $DFDR^{+}$  are used to construct equal-weight portfolios. The portfolios backtest technical rules in the stock markets in Africa, America, Asia, and Europe. The backtesting investigates the transaction costs and the IS and OOS profitability after transaction costs and risk-free rate. The same backtesting process also studies the persistence of the  $DFDR^{+}$  portfolios as well as the cross-validation of the constituent of the portfolios with a novel approach that preserves the natural order of the time series under study. This experiment structure makes the findings tangible for both academic and industrial audiences.

This Chapter's results explain why technical analysis is very popular in practice and highlight circumstances where technical analysis fails to generate positive performance. The null hypothesis is rejected for all twelve markets confirming that significant technical rules exist. This study shows that financial market dynamics play an important role in profitability and there is no evidence of one single profitable rule for trading over a long time. Emerging and frontier markets with a lower level of competition generally present better investment opportunities. No evidence is found that inclusion of fundamental factors like the financial stress index would increase the predictive power of technical analysis.

This study finds the missing element in the academic literature for reviewing technical trading: *rebalancing*. Technical trading focuses on detecting trends over short-term and speculates on these trends. The profitable rules are bound to a specific market and time. Testing decades of financial time series as in the literature (see among others, Cialenco and Protopapadakis, 2011; Fang *et al.*, 2014; Hsu *et al.*, 2016) is an impractical approach to analysing a trading strategy and therefore leads to criticism of technical analysis. The profitable trading strategies are short-lived and vanish once other market activists discover either an identical or a similar strategy. There are also cases where the portfolios failed to generate positive profit, however, there is always at least one profitable market over the ten years studied. The findings can be summarized as: *there is always a trout in the trading lake, but once caught it must be used while it is fresh.*

The DFDR<sup>+/-</sup> is applied to a trade analysis application. The empirical evidence shows its strength in comparing the large pool of candidate models. Such a pool of candidates exists across different fields e.g. finance, genetics, and computer science. One pathway for future research can be oriented toward unbiased comparison of alternative models in other disciplines by using the proposed procedure. This study compares a large pool of simple technical trading rules. With the rise of new AI models for trading, future studies in this area can be also oriented toward providing an unbiased comparison of complex AI models to find the most successful ones for trading.

## 5. Revisiting Financial Volatility Forecasting: Evidence from Discrete False Discovery Rate

### 5.1 Introduction

Market volatility is a “latent” economic variable which is not directly observable from the market. This is different from the price of a stock, which can be directly identified. Because market volatility is both latent and of great economic importance, a large and growing literature attempts to compare different volatility models based on their forecast accuracy. The magnitude of market volatility at each time is of crucial importance to option pricing models, quantitative risk management and asset pricing (Harvey and Whaley, 1992; Poon and Granger, 2003; Brooks and Persaud, 2003). Although volatility and risk have subtle differences, volatility is often used as a convenient proxy for risk in investment decisions. Christoffersen and Diebold (2000) argue that successful volatility forecasting improves such decisions. The aim of this study is to find the most successful volatility forecasting models through an unbiased simultaneous statistical comparison of all candidate models.

Various statistical models have been introduced to capture how volatility varies over time. The GARCH models of Engle (1982) and Bollerslev (1986) are widely recognized by both researchers and practitioners. Different extensions of the GARCH models have been introduced over time. The most common classes of GARCH models are: the absolute value model of Taylor (1986) and Schwert (1989), Exponential GARCH (EGARCH) of Nelson (1991), the asymmetric GJR-GARCH model of Glosten *et al.* (1993), the Threshold GARCH model (TGARCH) of Zakoian (1994), the Integrated GARCH (IGARCH) of Engle and Bollerslev (1986), Fractionally Integrated GARCH (FI-GARCH) of Baillie *et al.* (1996), and the RiskMetrics (RM) - a non-stationary extension- of JP Morgan (1996). Poon and Granger (2003) provide an extensive review of various classes of GARCH models.

The SV method of Taylor (1982) is an alternative technique for modelling time-varying volatility. SV focuses on the latent feature of conditional volatility and models it by a stochastic process (Sadorsky, 2005). Among others, Ghysels *et al.* (1996), Broto and Ruiz (2004), Sadorsky (2005), Asai *et al.* (2006), and Chan and Grant (2016) review the SV literature and estimation models. Yu (2002) and

Chan and Grant (2016) argue that SV models outperform their GARCH counterparts in the stock market and crude oil volatility forecasting.

A simple estimation of conditional volatility is the HAR by Corsi (2009) that utilizes daily, weekly and monthly components. The HAR provides an efficient long-memory regression model able to mimic the actions of different market participants (Hansen and Lunde, 2011). Despite its simple structure, HAR emerged as a very successful volatility forecasting method, often outperforming both GARCH and SV models (Corsi, 2009; Hansen and Lunde, 2011; Bollerslev and Quaedvlieg, 2016).

In this Chapter, I ask two research questions:

- 1) *Among the more popular volatility models, is there a specification or family of volatility models, that prevails in terms of forecasting accuracy? Particularly, are GARCH (1,1) and ARCH (1) truly the more accurate specifications?*
- 2) *Does the individual market characteristics play any role in that?*

To proceed, I need a measurement scale for the volatility models and a performance comparison framework to find the best ones. A practical scale to measure the true (latent) conditional variance is the Realized Volatility (RV), where daily RV is constructed upon aggregation of the squared high-frequency intra-daily returns (Andersen and Bollerslev, 1998; and Andersen *et al.*, 2003). RV is widely used in the literature (see among other Brooks and Persaud, 2003; Andersen *et al.*, 2005; Bollerslev *et al.*, 2016). As this Chapter's performance comparison framework, I use the MHT approach of  $\text{DFDR}^{+/-}$  discussed in Chapter 4.

In this Chapter's application, I review the literature in financial volatility forecasting, identify the most promising models and their variants, and I examine their accuracy through the 4<sup>th</sup> Chapter's discrete FDR approach. The volatility forecasting comparison literature can be traced back to Dimson and Marsh (1990). They study UK stock market volatility over the 1955-1989 and report that under an unbiased evaluation, "relatively sophisticated forecasting methods" underperform the naïve benchmarks. In contrary, Andersen and Bollerslev (1998)



study Deutschemark to US dollar (DM/\$) and Japanese yen to US dollar spot exchange rates over 1987-1992 and report that ARCH and GARCH models provide accurate volatility forecasts. Kang *et al.* (2009) study four GARCH classes (GARCH, IGARCH, CGARCH, AND FIGARCH) and use the forecast accuracy test (Diebold and Mariano, 1995) comparing one, two, and five days ahead forecasts for the three crude oil time series. Their results show a significant difference in forecasting OOS volatility between the competing oil price models. However, none of the mentioned papers uses a formal MHT error controlling approach to evaluate the models.

Wei *et al.* (2010) expanded the pool of Kang *et al.* (2009) to nine GARCH classes of models (RiskMetrics, GARCH, IGARCH, GJR, EGARCH, APARCH, FIGARCH, FIAPARCH and HYGARCH) and deploy the Hansen (2005) SPA test. Their results show that no model outperformed its counterparts in forecasting the volatility of Brent and West Texas Intermediate (WTI) crude oil prices. Bollerslev and Quaedvlieg (2016) introduce the HAR Quarticity (HARQ) model and use the RC test to compare their model against eight AR and HAR specifications. Their OOS study for the daily S&P 500 index RV shows limited significant differences between the candidate models. Hansen and Lunde (2005) bring together a large pool of three hundred and thirty volatility models and compare them based on the SPA test. They report mixed results for the FX market and the stock market. They find no significant improvement in favour of exotic extensions of GARCH models compared to a GARCH (1,1) and an ARCH (1) benchmarks for DM/\$ volatility. For IBM stock the benchmark is outperformed by at least some models in the pool. Nevertheless, their approach is not able to detect the superior ones individually. Bao *et al.* (2006) compare sixteen volatility models for Value at Risk (VaR), based on the RC test. They study risk models for three periods in five Asian countries. Their findings show that there are no superior models over the crisis periods and VaR models generally behave similarly. Esposito and Cummins (2016) expand the studied pool of Bao *et al.*, (2006) and deploy a  $k$ -FWER approach to MHT by using the Romano *et al.* (2008). Esposito and Cummins (2016) study sixteen OOS periods for one-step-ahead and ten-step-ahead forecasts and report that superior models exist under a more powerful MHT procedure. The literature of volatility forecasting has no record of MHT applications based on the FDR approach.

By considering alternative specifications for each class of volatility models and a range of different plausible parameterizations for each specification (see Hansen and Lunde, 2005), I construct a very large pool of possible volatility models. I study multiple classes of assets (FX, the stock market, and commodities) and conduct this Chapter's empirical analysis over five periods of one calendar year. One of the most common practices in volatility forecasting is to estimate a model based on historical data and project it forward (Figlewski, 1997; Bollerslev *et al.*, 2016; Esposito and Cummins, 2016). Analysis of future predictions can reflect the model's strength. Investigating one-day ahead is an almost unanimous practice for trading purposes among risk managers (Christoffersen and Diebold, 2000). Following this literature and the industry's well-accepted practice, I use the one-step-ahead RV forecasting accuracy as the performance measure. In terms of forecast evaluation, I compare the candidate models based on the novel DFDR<sup>+/−</sup> technique. First, I apply the stationary bootstrap resampling of Politis and Romano (1994) to generate the individual  $p$ -values. Then, Liang's (2016) adaptive approach is used to provide a more realistic analysis for the discrete space  $p$ -values. Next, the FDR<sup>+/−</sup> model of Barras *et al.* (2010) is deployed to control the expected probability of having a certain proportion of false positives. Finally, the step-wise algorithm in Bajgrowicz and Scaillet (2012) is used to find the set of true discoveries based on the FDR<sup>+</sup>. This procedure is introduced in Section 4.2 as Discrete FDR<sup>+</sup> (DFDR<sup>+</sup>). Thus, I try to find the set of superior volatility forecasting models by applying the DFDR<sup>+</sup> technique.

The rest of this study is organized as follows: in Section 5.2 the pool of 1,512 time-varying volatility models and their specifications are discussed. Section 5.3 focuses on the MHT method – DFDR<sup>+</sup> – used to compare the volatility models. Section 5.4 introduces the data applied and the performance metrics used to measure the accuracy of the individual models. Section 5.5 presents the empirical results and Section 5.6 provides some concluding remarks. Finally, several technical details and robustness checks are included in Appendix D.

## 5.2 The Conditional Volatility Pool

This Section covers the specifications of the time-varying volatility models considered in the forecast comparison practice. Given a daily price process  $\{x_t\}$ , the logarithmic return is given by  $r_t = \log(x_t) - \log(x_{t-1})$ ,  $t = 1, \dots, K + 1$ . The

dataset is then split into two sets of sub-samples: training (IS) and test (OOS). The first  $K$  observations are used as an IS set to estimate a predictive conditional volatility model for the last observation (OOS).

Given a  $\sigma$ -algebra  $\mathcal{F}_{t-1}$ , based on information available at time  $t - 1$ , the conditional density function for  $\{x_t\}$  is given by  $f(r|\mathcal{F}_{t-1})$ . The conditional mean and variance can be defined as  $\mu_{r_t} = E(r_t|\mathcal{F}_{t-1})$  and  $\sigma_t^2 = \text{var}(r_t|\mathcal{F}_{t-1})$  respectively. Following Hansen and Lunde (2005), I define the standardized return as  $\xi_t = (r_t - \mu_{r_t})/\sigma_t$  and the corresponding density function is given by  $g(\xi|\mathcal{F}_{t-1})$ . Given a parametric specification for  $f$ , the conditional mean and the variance of  $\{x_t\}$  are denoted by  $m_t = \mu(\mathcal{F}_{t-1}, \theta)$  and  $h_t^2 = \sigma^2(\mathcal{F}_{t-1}, \theta)$  respectively where  $\theta$  is a vector of parameters.

I construct my volatility forecasting models based on the  $\{\xi_t\}$  process. As a result, depending on the selection of the conditional mean and variance, multiple extensions of the standardized returns can be considered. The rest of this Section presents the characteristics of the candidate models studied for their predictive ability.

### 5.2.1 Conditional Mean and Variance

In order to generate the  $\{\xi_t\}$  series, an estimation of the conditional mean ( $\mu_{r_t}$ ) is necessary. I consider three estimations of  $\mu_{r_t}$  denoted by  $m_t$ . First, I assume a fixed zero-mean process  $m_t = 0$ . Second, I generalize my first case by allowing a non-zero mean based on the location of unconditional mean  $m_t = \omega_0$ . Finally, I consider a time-varying process based on the conditional variance as  $m_t = \omega_0 + \omega_1 \sigma_{t-1}^2$ . Financial time series are known to often have heavy tail distributions. To accommodate this, I consider four alternative distributions for the innovations, namely the Gaussian, Student's  $t$  (Bollerslev, 1987), skewed  $t$  (Hansen, 1994) and the Generalized Error Distribution (GED) (Nelson, 1991).

The RV derived from the intraday returns is used as a proxy for the latent  $\sigma_t^2$ . First, each trading day  $t$  is split into  $\mathcal{T}$  intraday periods, where the return for the  $t$  period is given by  $r_{t,t} = \log(x_{t,t}) - \log(x_{t,t-1})$  for  $t = 1, \dots, \mathcal{T}$ . Under the assumptions of zero-mean and conditionally uncorrelated returns for intraday periods, the daily RV is defined as:

$$RV_t = \sum_{t=1}^T r_{t,t}^2. \quad (5.1)$$

Most stock markets are open over certain weekday working hours and are closed on the weekends. In this case, only a subset of  $\mathcal{T}$  quotes is reported for each day, which explains the gaps that exist in the time series of the asset price. To account for this issue, two solutions are considered in the literature (Andersen and Bollerslev, 1998; Hansen and Lunde, 2005; and Huang *et al.*, 2013). The first solution is to scale the available information  $RV_t$  by a constant as a measure of volatility for the full day<sup>40</sup>. An alternative is to adjust the RV by adding the overnight return as:

$$ARV_t = RV_t + r_{t,on}^2. \quad (5.2)$$

In Eq. (5.2),  $r_{t,on}$  is the overnight return given by  $r_{t,co} = \log(x_{t+1,o}) - \log(x_{t,c})$ , where  $x_{t,c}$  is the closing price of day  $t$  and  $x_{t+1,o}$  is the opening price of the following day. Since Hansen and Lunde (2005) find that ARV is a ‘noisy’ measure for the daily volatility, I revisit it an alternative proxy to the  $\sigma^2$ . This allows me to identify possible differences between the models.

### 5.2.2 Forecasting Models

Four common families of volatility forecasting models are used in this study (GARCH, EWMA<sup>41</sup>, SV, and HAR). The specification of the GARCH family models is very common. The MA-based models, GARCH-MA and SV-MA are adopted from Chan and Grant (2016), where innovations are assumed to follow a first-order MA process. The asymmetric SV with leverage (SV-L) is adopted from Asai and McAleer (2011). Finally, HAR models are adopted from Corsi (2009) and log-HAR from Huang *et al.* (2013). The equations for each class of volatility models are given in Table 5.1.

[Table 5.1]

---

<sup>40</sup> I studied this approach as well. All estimated constants fall in a confined range of (0.9, 1.1) which is easy to capture by volatility forecasting models.

<sup>41</sup> RM is a specific case of the EWMA.

In GARCH  $(p, q)$  models (Eq.s (5.4 to 5.14)),  $p > 0, q > 0, a_u \geq 0, u = 1, \dots, q, \phi_r \geq 0, r = 1, \dots, p$ . For  $p = 0$ , the GARCH process reduces to ARCH  $(q)$  as in Eq. (5.3). In GJR-GARCH (Eq. (5.10)),  $I_{\{\varepsilon_t < 0\}}$  is a dummy variable accounting for leverage equal to the unit when the criterion  $\varepsilon_t < 0$  is satisfied or otherwise zero. In NGARCH and APARCH (Eq. (5.13 and 5.14)), the  $\delta$  is an extra parameter estimated along with other parameters. In asymmetric models (Eq.s (5.7), (5.8), (5.9), (5.10), (5.12), (5.14), and (5.17))  $\eta$  corresponds to the leverage effect. In EGARCH and SV-L (Eq.s (5.12 and 5.17)),  $E|e_{t-u}| = (\pi/2)^{-1/2}$  given  $e_t \sim \mathcal{N}(0, 1)$ . For other distributions (t-student, GED and skewed-t) the quantity should be computed accordingly as in Harvey and Sucarrat (2014). However, in this Chapter the same quantity  $(\pi/2)^{-1/2}$  is used as an approximation for other distributions, following Hansen and Lunde (2005).

In FIGARCH (Eq. (5.15)),  $a(L)$  and  $\phi(L)$  are lag operators such that  $a(L) = a_1 L^{(1)} + \dots + a_q L^{(q)}$  and  $\phi(L) = \phi_1 L^{(1)} + \dots + \phi_p L^{(p)}$ . The  $(1 - L)^d$  corresponds to fractional lag with degree  $d$  in the interval  $(0, 1)$ . Also,  $\zeta_t = \varepsilon_t^2 - \sigma_t^2$  is a zero-mean martingale process representing innovations for the conditional variance. In SV models,  $h_t = \log(\sigma_t^2)$  and  $\mu_h$  is the unconditional mean of  $h_t$ . The considered lag for all GARCH and SV models is one and two.

For EWMA models (Eq. (5.18)), a range of parameter  $v$  – interpreted as the decay factor – is studied from the set  $\{0.8, 0.82, \dots, 0.98\}$ . JP Morgan (1996) proposes a decay factor of 0.94 in their RM model. In HAR models (Eq.s (5.19 and 5.20)),  $\tilde{\sigma}_t^d$  is the daily conditional variance like  $\sigma_t^2$  in Eq.s (5.3 to 5.18). The  $\tilde{\sigma}_t^d$  is regressed on previous daily  $\tilde{\sigma}_{t-1}^d$ , weekly  $\tilde{\sigma}_{t-1}^w$  and monthly  $\tilde{\sigma}_{t-1}^{mn}$  values for the  $\{\tilde{\sigma}_t^d\}$  process (Eq.s (5.19 and 5.20)).

Based on the above, I generate 1,512 specifications by combining specifications for different variables. The candidate models are based on three specifications for conditional mean and two for conditional variance for generating the underlying standardized return series. Then, four innovation distributions are considered for the twenty classes of the forecasting models. Appendix D.1 includes more details regarding the above specifications and the characteristics of the models considered in the pool.

### 5.3 DFDR<sup>+</sup>

The DFDR<sup>+</sup> is a step-wise procedure to find the set of true rejections in MHT while controlling the false discoveries asymptotically. The procedure is able to efficiently balance the control of both Type I and Type II errors in model selection. This technique is built upon the premises of FDR, but it is adjusted for more adaptive hyperparameters and accounts for the discrete feature space. For the mathematical presentation of the DFDR<sup>+</sup> method, refer to Sections 4.2.1 to 4.2.3.

In this application, I use the DFDR<sup>+</sup> to find the set of outperforming volatility forecasting models. A continuous uniform distribution of the  $p$ -values may not be realistic, due to the finite number of the bootstrap replications and the dependence between alternative volatility models. Therefore, a discrete approach is more consistent with the studied universe of candidate models discussed in Section 5.2. The exact steps of the DFDR<sup>+</sup> approach that lead to the extraction of the superior volatility models are presented in Appendix D.2.

## 5.4 Model setup

### 5.4.1 Data

I conduct this Chapter's empirical analysis for six markets (three exchange rates, two stock indexes and one commodity) based on daily returns and high-frequency RVs. The exchange rates are European Union euro to US dollar (EUR/USD), British pound to US dollar (GBP/USD), and US dollar to Japanese yen (USD/JPY); stock market indexes are the DJIA and the Financial Times Stock Exchange 100 (FTSE 100); the commodity is the Gold spot price in US dollar (XAU/USD). The daily RVs are constructed as a summation of five-minute squared returns, following recent practices in volatility modelling (Li and Xi, 2016; Bollerslev *et al.*, 2016). The RV approximations are based on  $\mathcal{T} = 288$  for EUR/USD, GBU/USD, USD/JPY and XAU/USD,  $\mathcal{T} = 102$  for FTSE 100, and  $\mathcal{T} = 78$  for DJIA<sup>42</sup>. This Chapter's dataset starts on January 1<sup>st</sup>, 2012 and ends on December 31<sup>st</sup>, 2017. For each trading day,

---

<sup>42</sup> The exchange rates and the commodity are traded 24 hours on weekdays (Sunday to Friday, 22:00 to 22:00 GMT). The London Stock Exchange is open Monday to Friday, 08:00 to 16:30, while the New York Stock Exchange is open Monday to Friday, 14:30 to 21:00 GMT. The source of the applied data is Bloomberg.

one year of IS is used to predict the one-step-ahead conditional volatility. Thus, the OOS period is between January 1<sup>st</sup>, 2013 and December 31<sup>st</sup>, 2017. The OOS period is split into five sub-periods of one calendar year and the models are compared separately for each year to explore the dynamics in performance over time. Table 5.2 presents the summary statistics for the daily logarithmic returns of the time series.

[Table 5.2]

The dataset has more observations for globally traded securities (exchange rates and commodity) than the stock indexes. The mean and median return is negative for EUR/USD, GBP/USD, and XAU/USD. The standard deviation is higher in the gold market, while exchange rates appear to have lower standard deviations compared to the stock indexes. All securities are leptokurtic with negative skewness (except for EUR/USD). The statistic for the JB test rejects the null hypothesis of a normal distribution. This justifies the choice of the multiple innovation distributions under study. Ljung and Box (1978) statistics show rejection of serial independence for EUR/USD and FTSE 100 at 5% confidence level. The unit root tests of the ADF and the Philips and Perron (1988) (P-P) find no evidence for non-stationarity in the return series.

### 5.4.2 Performance Metrics

In this Section, I present a set of measures to compare the candidate models. In volatility forecasting literature it is widely common to use a loss function and a benchmark to compare the volatility models (Brooks and Persaud, 2003; Hansen and Lunde, 2005; Wei *et al.*, 2010; Huang *et al.*, 2013; Sermpinis *et al.*, 2015; Bollerslev *et al.*, 2016; and Wang *et al.*, 2018). Following the literature, I define  $\varphi_i$  as the test statistic for accuracy comparison as:

$$\varphi_i = -(\mathcal{L}_i - \mathcal{L}_0), i = 1, \dots, m \quad (5.21)$$

where  $\mathcal{L}_i$  is the calculated loss function for the candidate model  $i$  compared to the one of the benchmarks,  $\mathcal{L}_0$ . Finding the most appropriate loss function is a rather difficult task because there is no formal theory supporting such a selection (Lopez, 2001). I consider six loss functions constructed on the Mean Absolute Error (MAE), Mean Square Error (MSE), and logarithmic likelihood functions. The chosen

loss functions cover the ones in Hansen and Lunde (2005) and Wei *et al.* (2010). The specifications of the loss functions are presented in Table 5.3.

**[Table 5.3]**

The benchmark choice is crucial when using MHT approaches. In this Chapter's case, three different benchmarks are used to track the variations in the number of discoveries in each test. These benchmarks are an ARCH (1) as in Eq. (5.3), a GARCH (1,1) as in Eq. (5.4) with Gaussian innovations and RV as conditional variance, and the 90<sup>th</sup> percentile (PRC 90) of the entire pool based on the IS performance. Therefore, I use two model-dependent benchmarks and one pool-dependent benchmark. This choice allows the comparison between a model and the standard benchmarks in the literature and reveals whether there are any significant differences between the top-performing models. By using the proposed setting, I transform the loss functions into the gain functions to find the models on the right (positive) tail. The null hypothesis is then redefined to test whether a candidate is able to provide higher accuracy compared to the alternative benchmarks of ARCH (1), GARCH (1,1), or PRC 90.

## 5.5 Results

In this Section, I present the outcomes of the empirical analysis. To find the set of superior volatility models, five years of one-day-ahead predictions were generated. Then a test statistic was calculated for each year based on a chosen loss function and compared to a benchmark by the DFDR<sup>+</sup>. In the testing procedure, I set the FDR controlling target to 10%. I compare the models in the volatility pool based on their performance over each of the five study periods separately (2013 to 2017). The results presented in the rest of this Section are based on the  $MSE_1$  as the loss function. Appendices D.3 to D.8 present the same type of analysis for alternative choices of the loss function.

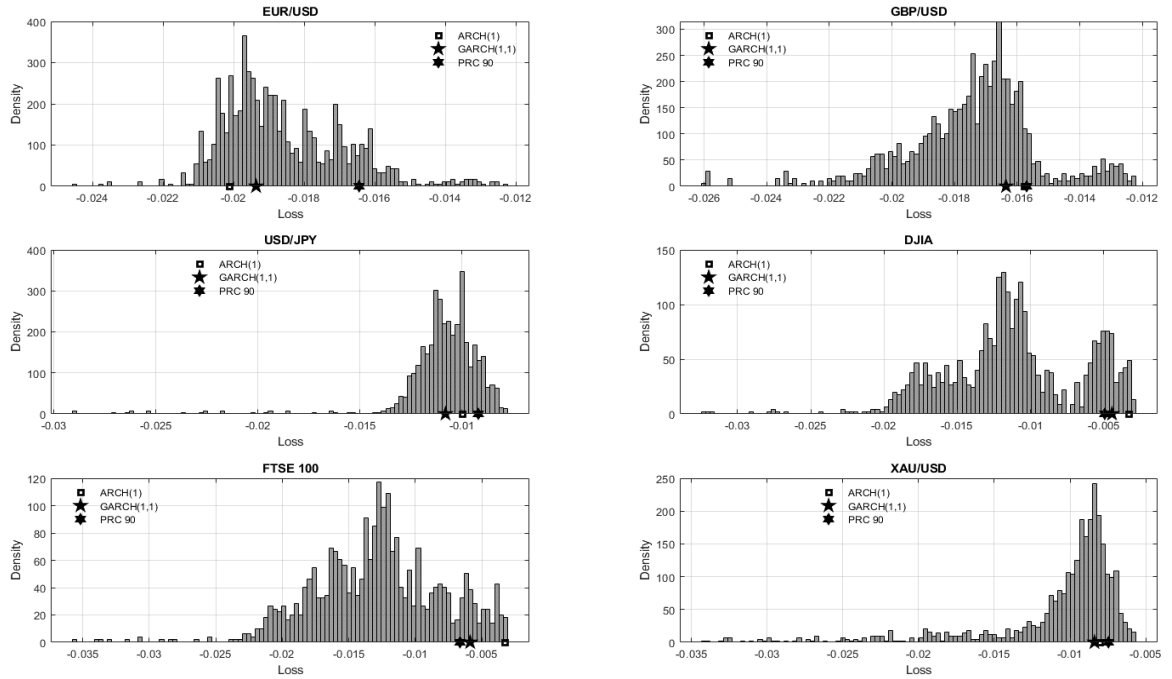
### 5.5.1 Model Performance

The first step in evaluating the pool's performance is to depict the population density for accuracy of the volatility models. The aim is to examine the



characteristics of the distribution and identify potential points of increased density. This is shown in Figure 5.1 below:

**Figure 5.1: Model Performance Density**



**Note:** The figures present the density of average loss across the pool based on the  $MSE_1$  benchmark. The  $x$  axis is  $\tilde{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  stands for the OOS subperiods of one year. The polygonal presents the location of the benchmarks studied. The square, pentagram, and hexagram correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

From Figure 5.1, I note that the population distribution is characterized by its wide range, sparsity and clear evidence of increased density around certain points. The wide range is due to a small proportion of the outlying models. The sparsity and increased density are due to the dependence structure in the volatility pool and the modest changes between similar models. The sparse pattern of the sample loss justifies the choice of a discrete  $p$ -value assumption and the DRB method used for the FDR modelling. The performance of the three benchmarks is highlighted. The ARCH (1) and the GARCH (1,1) loss are mostly either overlapping or very close. However, their positions are relatively dynamic with respect to the whole pool. In the currency and commodity markets, the ARCH (1) and GARCH (1,1) are not in the top 10 percentiles, while for the stock market they outperform 90% of the models in the pool. Therefore, volatility models come with a range of performances and compete in a close race without a clear winner. Appendix D.3 presents similar results for the other loss functions. The comparison of the loss functions shows that the sparsity pattern persists but varies remarkably

from one market to another. In the following sections, the MHT is used to find the superior models.

### 5.5.2 True Discoveries

Table 5.4 presents the number of outperforming models for different markets over the OOS period (2013 to 2017). In the table, the volatility models are compared against the ARCH (1), the GARCH (1,1), and the PRC 90 benchmarks. The values correspond to the size of the true discoveries set measured for each calendar year.

[Table 5.4]

There are several interesting findings derived from the table. Initially, the various specifications of the volatility models provide a significant edge in forecasting over the benchmarks. The PRC 90 is the most difficult benchmark to beat on average for all assets. The set of superior volatility models are time and asset dependent. For all markets, there is at least one case where no significant difference in the volatility pool is detected. The relative strength of the ARCH (1) and GARCH (1,1) models vary based on the market and the loss functions. In FX and commodity markets, GARCH (1,1) is a more difficult model to beat. This is inferred from the multiple cases where the set of discoveries is confined to the model with lowest  $p$ -value. In the stock markets the ARCH (1) has a higher strength compared to both GARCH (1,1) and the PRC 90 as the number of models beating the ARCH (1) is lower on average compared to the counterpart benchmarks. Finally, by comparing all markets it is observed that the average number of rejections is the least with the DJIA and greatest with the EUR/USD. Appendix D.4 explores the same analysis for other loss functions. Among the loss functions, QLIKE and  $R^2\text{LOG}$  give the lowest and highest average number of rejections respectively.

### 5.5.3 Distribution

In this Section, I investigate the role of innovations assumed for the volatility models. GARCH and SV models require a distribution for innovations, whereas EWMA and HAR models do not use it. I consider four distributions for the GARCH models and three (Gaussian,  $t$ , and skewed  $t$ ) for the SV models. Therefore, there

are 372 volatility models with Gaussian,  $t$ , and skewed  $t$  distributions, 324 with GED, and 72 with no distribution. Table 5.5 exhibits the average proportion of the models from each distribution found to be significant in each test.

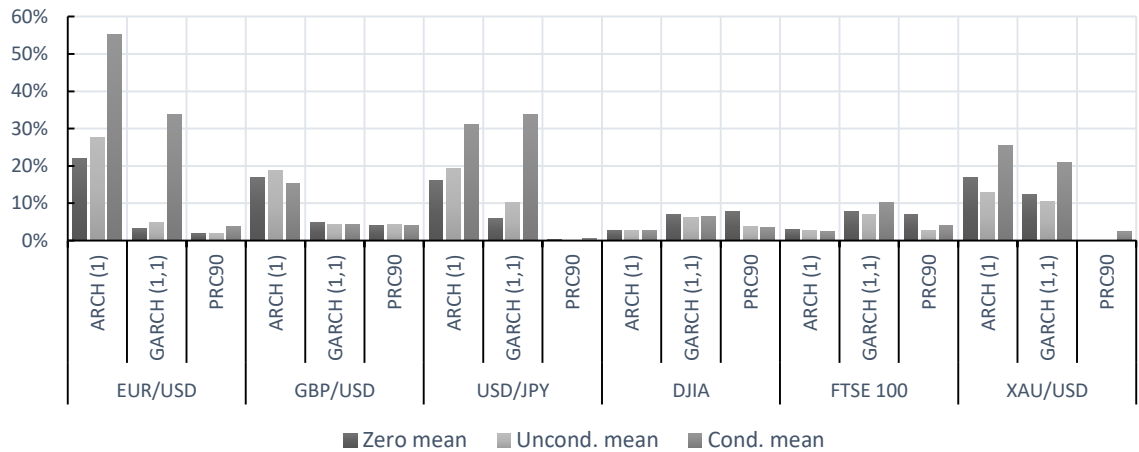
[Table 5.5]

The results show that the innovations are crucial to volatility modelling. A very limited number of models without a fitting distribution are able to outperform the benchmarks and they never end up in the top 10 percentiles of the pool. For the exchange rates (EUR/USD, GBP/USD, and USD/JPY) the GED performed the best on average. For the DJIA, the  $t$  is the leading distribution whereas, for the FTSE 100 and the Gold, the skewed  $t$  has the highest survival rate. Among the examined distributions, the skewed  $t$  forms the highest percentage of significant models across the markets. Inspection of alternative loss functions in Appendix D.5 shows that the best distribution is unchanged with the QLIKE loss function, but changes to the GED for the MAE functions. On the other hand, the  $R^2\text{LOG}$  shows the best survival rates for the Gaussian distribution. A paired  $t$ -test for equal means based on all loss functions, study periods and markets, is conducted to compare the average proportions of different distributions to the Gaussian. The category of models without any distribution is the only significantly different one.

#### 5.5.4 Mean Estimation

The standardized return is built on two variables: a mean and a variance component. Here, I explore the contribution of the mean estimation to the superior volatility models. I study three different specifications to estimate the daily return for generating standardized return series in Section 5.2. The specifications are a zero-mean, an unconditional mean and a time-varying (conditional) mean regressed on the previous day RV or ARV. The number of models with each specification is equal to 504. Figure 5.2 plots the average proportion of surviving models based on each estimation technique.

Figure 5.2: Survival Rate for Alternative Means Specifications Across the Markets.



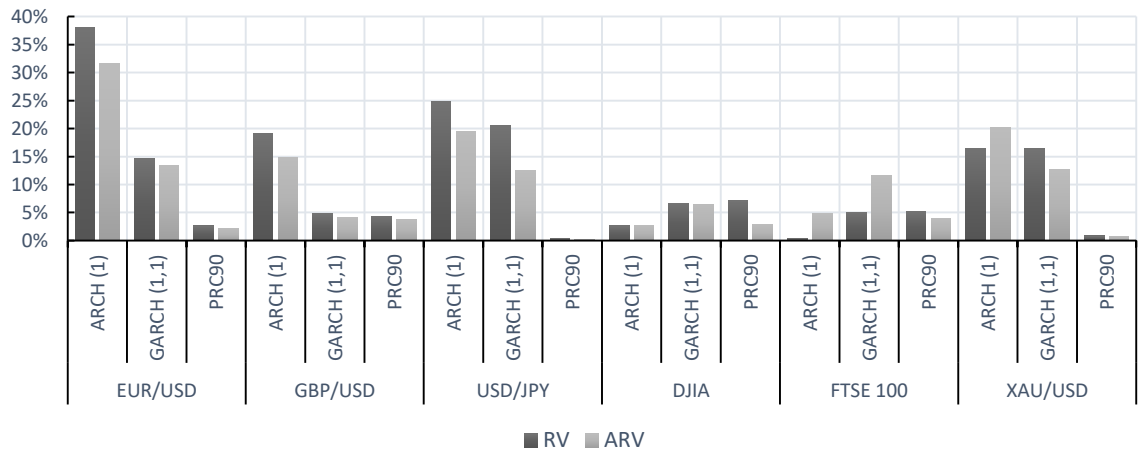
**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_1$ .

The results show an increase in the proportion of surviving models as the accuracy of the estimated location is improved. The paired  $t$ -test shows no difference between the average for zero-mean and unconditional mean processes, while the survival rate of the conditional mean is significantly different from both other processes. The order of the mean specifications is the same for all loss functions exhibited in Appendix D.6.

### 5.5.5 Conditional Variance

The second component forming the standardized return is the variance proxy. In Section 5.4, I focus on the role of the mean estimation, whereas in this section I focus on the conditional variance. I use two alternative proxies: the RV (Eq. (5.1)) and the ARV (Eq. (5.2)). Half of the models in the pool (756) are allocated to each specification. Figure 5.3 demonstrates the average success rates for models with RV and ARV.

Figure 5.3: Conditional Variance Survival Proportion Dynamics Across the Markets



**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five-minute squared returns as in Eq.s (5.1 and 5.2) respectively. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_1$ .

Unlike with the conditional mean, there is not a solid consensus about the variance specifications. The paired  $t$ -test shows significant differences between the two variance specifications. For the PRC 90, the RV is the dominant variance proxy across the markets on average for all loss functions. Additionally, the RV leads to a higher (or almost similar) rate of surviving models in the currency market for all three benchmarks. For the DJIA and Gold, the ARV performs better for the ARCH (1). For the UK stock market, ARV is the leading variance for both ARCH (1) and GARCH (1,1) benchmarks. The alternative loss functions are demonstrated in Appendix D.7. The patterns are generally stable and in favour of RV for the different loss functions except for the  $MAE_2$  where the ARV performs better on average across the markets. The relative difference between specifications is once again at a maximum with the  $R^2LOG$ .

### 5.5.6 Class

In this Section, I try to find the successful classes of volatility forecasting models. I study fourteen classes of GARCH models, three classes of SV models, two classes of HAR models and one class of EWMA models (see Table 5.1 for equations and Appendix D.1 for the count of models in each class). Table 5.6 displays the average proportion of surviving models in each class based on all three benchmarks.

[Table 5.6]

The results show that the best performing class of models across the exchange rate and commodity markets is by far the IGARCH. The gap between the IGARCH and second top performing class (NGARCH) is maximized to over 50% for GBP/USD. The SV and the SV-MA are the top two classes for both US and UK stock markets. The highest average across the markets belongs to the IGARCH and the SV. The top-class patterns are subject to the choice of loss function. The dynamics across loss functions are presented in Appendix D.8. For the MAE scales, the only difference is that ARCH outperforms the SV for the FTSE 100. The QLIKE changes the top-performing class for the Gold to the FI-GARCH. The major divergence in patterns comes with the  $R^2\text{LOG}$  where the leading class of volatility models is the SV with leverage (SV-L) for the currency markets and on average. The TGARCH, GJR-GARCH, and IGARCH are the best fitting forecasting models for the DJIA, FTSE 100 and Gold respectively.

Four classes of models – log-GARCH, EGARCH, HAR, and log-HAR – fail to beat the benchmarks in any markets based on the  $\text{MSE}_1$ . These low-performing models account for the outliers in Figure 5.1. The underperforming pattern is consistent with most loss functions. However, with the  $R^2\text{LOG}$  even the least successful classes manage to beat the benchmark at least for some periods. This is the only loss function where the leveraged models perform superior to their counterparts.

## 5.6 Conclusion

This study performs a statistical comparison of 1,512 models to find the most accurate ones in forecasting the OOS volatility. The study is conducted for the most liquid assets in financial markets: EUR/USD, GBP/USD, and USD/JPY from currency pairs, DJIA and FTSE 100 from stock indexes, and the Gold from the commodity market. A novel bootstrap MHT procedure ( $\text{DFDR}^+$ ) is applied to compare the pool of volatility models based on six loss functions. The test hypothesis is that all volatility models are equally accurate compared to the most common benchmarks in the literature.

The results suggest that, if the models are specified correctly they are able to beat the ARCH (1), and GARCH (1,1) benchmarks, and 90% of their counterparts in one-step-ahead forecasting accuracy. I study the characteristics of the set of best forecasting models from 2013 to 2017. The range for the number of

outperforming models is from zero to several hundred based on multiple factors. The factors are time, market, and loss functions. The superior models are generally constructed with a fat-tailed distribution and a continuously updated estimation of the return series mean.

Several classes of GARCH and SV models consistently outperformed the benchmarks with a significant lead compared to the other classes. Differential models like IGARCH and standard SV models exhibited the highest success. Some classes consistently failed to beat any benchmarks across the markets. The logarithmic models (log-GARCH and EGARCH) and those without a fitting distribution (RM and HAR) exhibit the least success. The latter is in contrary to the apparent success of HAR models in the volatility forecasting literature (see among others Corsi, 2009; Bollerslev and Quaedvlieg, 2016). However, none of the previous works presented any formalized hypothesis testing results for HAR family relative to GARCH or ARCH benchmarks. Therefore, one major finding of this research is that under an unbiased evaluation, HAR models fail to produce significant improvements over this Chapter's benchmarks.

The findings in this Chapter put forward the importance of the MHT framework in risk management tasks, such as finding the best models for financial forecasting and decision making. Although there is no single "sacred" model to predict the future under market certainty, the MHT helps to find the most promising models statistically proven to be better than many other traditional approaches. New research in this field could explore new MHT techniques with a higher power, alternative proxies of volatility, exploration of the optimal loss functions for each market, and coupling the time-related characteristics of the discoveries with the dynamics in the markets or the economic fundamentals.

## 6. Conclusion

### 6.1 Summary

This thesis introduces several quantitative solutions for investments through four empirical chapters. The first chapter introduces an expert system paradigm (Figure 1.1) for decision-making. The paradigm first generates a pool of predictors from the raw dataset. The raw dataset is determined by the underlying problem. The predictors can be simple technical trading rules, AR models, or even betting odds. In the next step, the most informative predictors are chosen from the potential pool of inputs by an SI method. The final step in the paradigm is allocating optimal weights to the selected predictors to make predictions.

The second chapter provides the first application of the proposed paradigm to FX. A pool of eight thousand technical trading rules is used to generate profitable trading strategies. This chapter combines an FWER approach to hypothesis testing with four Bayesian AI methods – NB, RVM, DMA/DMS, and BNN. The results show the success of the proposed system in predicting the most liquid currency pairs.

The third chapter introduces an original AI method, namely CF. The CF is accommodated in the Figure 1.1 framework and is combined with a probabilistic SI component. The CF is then applied in a sport-betting application, where poor predictions may lead to total loss of the investment. The performance of CF is compared to three benchmarks: OP, RVM and ANFIS. The results in the third chapter show the superior performance of the CF compared to the other models and serve as further evidence for the success of the proposed paradigm.

The fourth chapter develops a new tool for the Filter layer of Figure 1.1. The DFDR<sup>+/-</sup> approach can compare many structurally similar models at once. The new method is validated by Monte Carlo simulations and then applied in analysing twenty-one thousand technical rules. The results show that technical trading can be profitable if frequently updated investment strategies are used with the latest market developments.



The fifth chapter shows another possible application of the proposed MHT approach. The DFDR<sup>+/·</sup> approach is used to compare one thousand five hundred volatility-forecasting models. This chapter focuses on illustrating the properties of the most successful models for forecasting one-step-ahead volatility. This chapter's findings show that some classes of volatility models can significantly outperform the benchmarks ARCH (1), GARCH (1,1), and 90% of their counterparts.

The results from the four empirical chapters can be summarized as: *quantitative methods are practical and productive*. The proposed tools in this thesis, together with the unceasing development of statistical methods, and exponential increase in computer capacity, are an alarming combination. Quantitative trading does not require any finance knowledge and is still able to predict markets, meaning that AI may replace experienced traders. The finance literature needs more focus on OR to keep up with the developments in other fields. Academics should also be aware of differences in theory and in real-world practice. The industry puts a lot more weight on what can increase the shareholders' profit irrespective of the theories. This thesis tries to compare statistical methods from an investor point of view to bridge that gap. Studies like this should help to keep the academic finance literature closer to developments in top-tier trading desks. Pursuing this path in future theses in quantitative finance can safeguard against divergence between academia and industry.

## 6.2 Limitations

This thesis tries to propose alternative systems for financial decision-making. The proposed systems rely on inference and/or learning from the developments observed in the datasets. This replicated experience-based knowledge offers major improvements in investment decisions; however, it has certain constraints of which two main ones are addressed here.

The first and foremost issue with the AI models is the unseen observations. This corresponds to cases where the OOS is very different from the IS. In this situation, the predictive model either does not have any information about the new observations or has irrelevant information with regards to new developments in the OOS. For instance, consider the stock trading application in Chapter 4. Since no information about the global financial crisis over 2008-2009 is provided to the

predictive models the performance around this time is extremely volatile. Humans naturally have the inherent feature of instinct for such complex situations. The existing inference and learning paradigms are unable to replicate the instinct and offer accurate decisions based on that. Nevertheless, finding a solution to this problem could potentially lead to machines that are both more flexible and unbiased in cognitive behaviour.

The second problem confining this research is computing resources. In the era of big data and data-science finding the best strategies based on the countless sources of information seems very promising. However, analysing such datasets can be a difficult task from a computation point of view. The Bayesian models in Chapter 2 (BNN, DMA, and DMS) provide excellent performance by learning from 15 explanatory variables. However, increasing the number of inputs from 15 to 20 slows down the computations 32 times. Another computational limit encountered in this research was in Chapter 4. The Bootstrapping procedure used there requires the generation of random orders 1000 times for matrices of 21000 rules by 500 periods. In addition, Chapter 5 retrains 1500 volatility-forecasting models every day for more than 1000 days. None of these computations was feasible on a standard PC. To deal with the dimensionality of empirical studies, cloud- and parallel- computing was used. Further robustness checks and benchmark comparison could be explored if more computing resources were available. With the growth of computers, less of these problems are expected in future.

### 6.3 Future Works

The focus of this thesis was to combine AI and SI approaches to articulate decision systems for investment management. I tried to quantify the marginal contribution of the new models compared the common benchmarks within the borders of the proposed paradigm presented in Figure 1.1. Improvements to the discussed methods can be made both external and internal to the boundaries of the paradigm. Externally, data-science is rich in approaches toward designing predictive models. Thus, one area of interest is to see how different methods for predictive systems perform compared relative to each other in practice. Internally, the proposed paradigm is composed of three main components: a pool of alternatives, an AI model and an SI one. All these components could be developed further in future.

The academic literature is suggesting new predictive models in every heartbeat. As time passes, the datasets become more accessible and accurate. Combining the new techniques with the new datasets leads to new applications with larger pools of candidates.

The future works in AI can be oriented toward using optimization techniques in Bayesian models to speed up the learning procedure. This allows maximizing the benefits of strong learning capacity of the Bayesian models while controlling the computational burden. Also analyzing different AI systems through an MHT procedure can show the true differences in the predictive ability of candidate models. The application for such research can be any field where pattern recognition is important e.g. finance, business analytics, engineering. Another area of interest could be modifying ML algorithms to form a conditional approach similar to CF. Greater performance compared to the unconditional model is expected since this approach can reduce the exposure of the model to the unmatched observations in the OOS.

The MHT plays an integral role in three out of four essays of this thesis. It can provide an understanding of which subset of candidates is able to perform superior relative to a benchmark under an unbiased evaluation. The Monte Carlo simulation in Appendix C shows the major improvement from an early testing procedure to the most recent ones like  $DFDR^{+/-}$ . Nevertheless, the power of the MHT procedures yet needs to be improved. For instance, the  $DFDR^{+/-}$  is able to find only half of the true discoveries when performances are close among the candidates (see Table C.2). Consequently, the unnecessary conservativeness of the MHT procedures is still an issue has to be resolved in future.

## Appendices

### Appendix A

#### A.1 Technical Trading Rules

Technical trading strategies involve using quotes for open, high, low and close prices along with the trading volumes of every ticker under study. Their purpose is to recognize patterns or trends in the price charts. I consider five classes of technical indicators. The studied classes are FIR, MA, S&R, CB, and OBV. Short descriptions of these rules are presented in the following subsections.

##### A.1.1 FIR

The filter strategy is based on making a financial decision to undertake a long or short position when the security price moves a certain amount e.g.  $x$  percent upward or downward. A buy order is placed when the  $x$  percentage upward movement is seen in the market and this position is held until the price falls  $x$  percent where the position is first neutralized. Then a short position is opened and kept until a subsequent upward movement is seen. Movements that are smaller than the filter level in either direction are discarded as noise.

Tuning the FIRs is at the analyst's discretion. The definition of upward/downward movement, the holding process and liquidating/closing the position can be subjective. Upward (downward) movements are recognized as an uptrend when the price exceeds the last high (low) by  $x$  percent. The last high (low) can be defined either as the highest (lowest) close price observed in a long (short) position; or the maximum (minimum) close price over the last  $d$  days. The position holding can also be modified. Another case is considered in which the position is opened by the FIR and held for  $h$  days where signals over this period are ignored. The strategy may also include a neutral position where positions are closed in case of  $y$  percent backward movement compared to extrema level. The filter level for liquidating the position must be less than the filter level for opening a position.

Consider the following sets of possible  $x, y$ :

$x \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30, 40, 50\}$  in %

$y \in \{0.5, 1, 1.5, 2, 2.5, 3, 4, 5, 7.5, 10, 15, 20\}$  in %,

$$(\#x = 24, \#y = 12)$$

Then, the number of  $x - y$  combinations, given that  $y < x$ , are  $\#(x - y) = 185$ .  
For these combinations, I experiment with the following  $d, h$ :

$$d \in \{1, 2, 3, 4, 5, 10, 15, 20\}, (\#d = 8)$$

$$h \in \{5, 10, 25, 50\}, (\#h = 4)$$

Based on the above, in this application I examine a total of 497 filter rules as calculated below:

$$\#F = \#x + \#x \times \#d + \#x \times \#h + \#(x - y) = 24 + 192 + 96 + 185 = 497 \quad (\text{A.1})$$

### A.1.2 MA

In technical trading MAs play an integral role. Trends are not considered robust until they are reflected in MAs. When a change in price is also visible in the MAs, it means that the news or change source is important enough to last over a period of time and can be taken into account. Uptrends start to form when a fast MA exceeds the slow MA. The fast MA can simply be the price quote or a short-term average. The long positions are kept so long as the price remains above the MA benchmark. When the price falls below the MA, the downtrend is initiated. At this point the previous position is liquidated and a sell position is opened. The new position remains open until another upward penetration is observed. MA crossover-based strategies may appear from a wide variation. In second chapter, simple forms of MA crossover along with some filter and delays are taken into account. A common application of MA is having a fast and slow MA and looking for their crossovers signalling/which signal up and downtrends.

A buy signal can be generated when the fast MA goes beyond the slow MA. The fast and slow MAs come with parameters  $n$  and  $m$  respectively showing the number of days taken into consideration ( $n < m$ ). Similarly, a sell signal is created

when the fast MA drops below the slow MA, which suggests the formation of a downtrend. The fast and slow MA strategy can be accompanied by a band ( $b$ ) filter to avoid the noise in trend detection. Trends are deemed solid only if the fast MA can exceed the slow MA by  $b$  percent. Alternatively, the time lag  $l$  is considered between opening a position and taking any action. During the lag period, all signals are ignored. Another innovation is holding each position for a fixed period of  $h$  days no matter what the signals are after the opening of the positions. In second chapter's application, I consider innovations separately and impose only one filter at a time.

I consider the following sets of possible  $n, m$ :

$$n \in \{2, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200, 250\}, (\#n = 15)$$

$$m \in \{2, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200\}, (\#m = 14)$$

Then, the number of  $n - m$  combinations, given are  $m < n$  are  $\#(n - m) = 105$ .

For these combinations, I experiment with the following  $b, l, h$ :

$$b \in \{0.1, 0.5, 1, 1.5, 2, 3, 4, 5\} \text{ in } \%, (\#b = 8)$$

$$l \in \{2, 3, 4, 5\}, (\#l = 4)$$

$$h \in \{5, 10, 25, 50\}, (\#h = 4)$$

The band filter is set at 1% . A 10-day holding period is applied to all combinations of MA crossovers. For the fast and slow MA I set respectively  $n = 1, 2, 5$  and  $m = 50, 150, 200$ . I also include 9 cases of double-filters. Based on the above, in this application I examine in total 2049 MA rules as calculated below:

$$\begin{aligned} \#MA &= \#n + \#(n - m) + \#b \times (\#n + \#(n - m)) + \#l \times (\#n + \#(n - m)) + \#h \times \\ &(\#n + \#(n - m)) + 9 = 15 + 105 + 960 + 480 + 480 + 9 = 2049 \end{aligned} \quad (\text{A.2})$$

### A.1.3 S&R

The S&R trading rules are based on the premise that the price should remain in a trading range capped by a resistance and floored by a support level. Breaching these levels suggests that a stock or an exchange rate would move in the same direction. The S&R rules are constructed similarly to the FIRs. The only difference is that trading signals are generated when the rate under study breaks the support or resistance barriers by a certain percentage. The S&R levels can be defined as the intra-day low and intra-day high quotes over the past  $n$  days. Another variation in the definition of the S&R is to calculate the support and resistance level based on the minimum and maximum closing prices over the past  $e$  days. Alternative S&Rs are set by using a fixed band filter for noise removal: the holding period  $h$ , the  $l$ -day lag before making any decisions, a combination of a fixed holding period on a position, and a delay in decision making before undergoing any new positions.

Based on the above, I consider the following possible sets:

$$n \in \{5, 10, 15, 20, 25, 50, 100, 150, 200, 250\}, (\#n = 10)$$

$$e \in \{2, 3, 4, 5, 10, 20, 25, 50, 100, 200\}, (\#e = 10)$$

$$b \in \{0.1, 0.5, 1, 1.5, 2, 3, 4, 5\} \text{ in } \%, (\#b = 8)$$

$$l \in \{2, 3, 4, 5\}, (\#l = 4)$$

$$h \in \{5, 10, 25, 50\}, (\#h = 4)$$

In accordance with these sets, I examine a total of 1220 S&R rules as calculated below:

$$\begin{aligned} \#S\&R &= [(\#n + \#e) \times (1 + \#h)] + [(\#n + \#e) \times (1 + \#h) \times \#b] + [(\#n + \#e) \times \#h \times \#l] \\ &= 100 + 800 + 320 = 1220 \end{aligned} \quad (\text{A.3})$$

### A.1.4 CB

Based on the principles of S&R, practitioners can detect time-varying support and resistance levels that drift together within a certain range. This creates the so-

called trading channel. Once a trading channel is formed, then a CB rule can be applied. The premise behind the CB rule is that once the trading channel is breached, there will be a substantial trend towards the same direction. A channel is formed when the highest observed price remains within a  $c\%$  range above the lowest price over the past  $n$  days. The trend is considered significant, when the price breaks one of the channel borders, which generates a buy (sell) order after an upward (downward) breakout. As in the previous categories discussed in Sections A.1.1 to A.1.3, I also consider CB alternatives with a fixed filter band  $b$  and holding period  $h$ .

Here I look at the following possible sets:

$$n \in \{5, 10, 15, 20, 25, 50, 100, 150, 200, 250\}, (\#n = 10)$$

$$c \in \{0.5, 1, 2, 3, 5, 7.5, 10, 15\} \text{ in } \%, (\#c = 8)$$

$$b \in \{0.1, 0.5, 1, 1.5, 2, 3, 4, 5\} \text{ in } \%, (\#b = 8)$$

$$h \in \{5, 10, 25, 50\}, (\#h = 4)$$

Given  $b < c$ , the number of  $c - b$  combinations are  $\#(c - b) = 43$ .

In this application I examine in total 2040 CB rules:

$$\#CB = \#n \times \#c \times \#h + \#n \times \#(c - b) \times \#h = 320 + 1720 = 2040 \quad (\text{A.4})$$

### A.1.5 OBV

In the technical trading context, prices and trading volumes are expected to move together. Trading volumes confirm the potential significance of price moves. In case of major economic events or important news, increased trading volumes reflect decisions in favour of or against the price change. Therefore, monitoring the volumes and their changes can be a useful source of information for the practitioner. The OBV line is simply a running total of positive and negative volumes. In other words, if the closing price is above (below) the prior close price, then the current OBV is the sum (difference) of the previous OBV and the current



volume. When the volume is not increasing during bullish days, it is a sign that buying pressure is weakening and the upward trend is probably not sustainable. OBVs are usually used with MAs to generate trading signals. In this scenario, the average OBV is calculated and then combined with slow and fast MAs. In second chapter's application, I use the MAs as in Section A.1.2, excluding the 9 double-filter cases. Based on these, I examine a total of 2040 OBV rules as calculated below:

$$\#OBV = \#n + \#(n - m) + \#b \times (\#n + \#(n - m)) + \#l \times (\#n + \#(n - m)) + \#h \times (\#n + \#(n - m)) = 15 + 105 + 960 + 480 + 480 = 2040 \quad (\text{A.5})$$

### A.1.6 Trading Universe

The Trading Universe (*TU*) consists of the total number of trading rules reported in the previous subsections:

$$\#TU = \#F + \#MA + \#S\&R + \#CB + \#OBV = 497 + 2049 + 1220 + 2040 + 2040 = 7846 \quad (\text{A.6})$$

## A.2 Sharpe Ratio

Tables A.1 to A.3 present the trading performance of all combinations in terms of annualized Sharpe ratio after transaction costs.

[Tables A.1 to A.3]

## Appendix B

### B.1 CF Illustration Example

In this Appendix, I present a detail CF example for the Premiership, the first forecasting exercise and the game result as a target. The IS is the seasons 2006 to 2007, 2007 to 2008 and 2008 to 2009 and the OOS is the season 2009 to 2010.

As mentioned above, the first step is to collect the generated RVs when feeding all the inputs of Table 3.1 into RVM. For the specific case, a set of 12 RVs are selected and based on these, the RVM generates a series of IS forecasts. The RV set is presented below in Table B.1.

[Table B.1]

Based on these forecasts, a set of FRs is derived. In this exercise, 13 rules are generated which cover the whole IS. The rules have the form:

*If (BbAv>2.5 is cluster  $A_{i,1}$ ) and (BbAHh is cluster  $A_{i,2}$ ) and (BbAvAHH is cluster  $A_{i,3}$ ) and (BbAvAHA is cluster  $A_{i,4}$ ) and (PtH3H is cluster  $A_{i,5}$ ) and (PtA3A is cluster  $A_{i,6}$ ) and (StH1 is cluster  $A_{i,7}$ ) and (StA1 is cluster  $A_{i,8}$ ) and (CkH3 is cluster  $A_{i,9}$ ) and (CkH2 is cluster  $A_{i,10}$ ) and (CkH1H is cluster  $A_{i,11}$ ) and (CkA2A is cluster  $A_{i,12}$ ), then the output is the result of regression  $\delta$ .*

where  $A_{i,k}$  is the cluster specified for input element  $k$  of the  $i$ -th rule. The FRs are specified by the premise and consequent parameters. The premise parameters determine the clusters specification: the centres ( $c_i$ ) and the standard deviations ( $\sigma_i$ ) for each rule. These are presented in Table B.2.

[Table B.2]

The output of each rule is associated with a regression  $\delta$  specified with consequent parameters. The parameters of these regressions for each rule are presented in Table B.3.

[Table B.3]

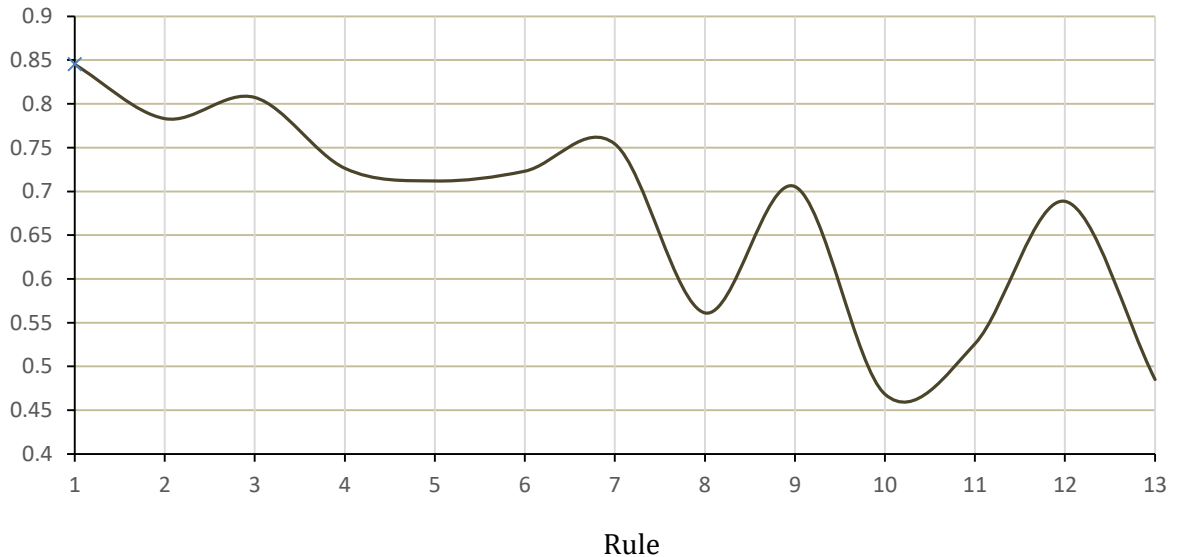
From the table, it is easy to extract the regression specification for the first rule  $\delta_1^*$  :

$$\begin{aligned} \delta_1^* = & -2.706 + 1.042x_{1,1}^* - 0.755x_{1,2}^* + 1.0721x_{1,3}^* + 0.484x_{1,4}^* - 0.041x_{1,5}^* + \\ & 0.207x_{1,6}^* - 0.138x_{1,7}^* - 0.088x_{1,8}^* - 0.197x_{1,9}^* - 0.016x_{1,10}^* + 0.152x_{1,11}^* + 0.074x_{1,12}^* \end{aligned} \quad (B.1)$$

where  $x^* = [x_1^*, \dots, x_{12}^*]$  is the vector of observed values for  $K = 12$  relevance vectors.

The next step is to evaluate the strength of the generated rules. To determine the CF threshold, the algorithm of Section 3.2.2.2 is followed. Firstly, all rules are evaluated based on the Gaussian membership function. For each of these rules, the average firing strength ( $w^{1/12}$ ) is estimated. The FRs are sorted based on the average firing strength over the IS in a descending order (the sorted list). The 90<sup>th</sup> percentile of the sorted list determines the endogenous threshold. In this case, the rule number one is the 90<sup>th</sup> percentile of the sorted list ( $\lambda = 1$ ). Thus, the endogenous threshold level is set to  $w_1^{1/12} = 0.85$ . Figure B.1 presents the average firing strength of all 13 rules.

**Figure B.1: Average Firing Strength for the  $w^{1/12}$**



**Note:** The figure presents the average firing strength for each one of the 13 generated rules based on the 12 selected RVs.

As the general threshold level (0.90) is greater than the endogenous one (0.85), the effective threshold ( $\theta$ ) is set to 0.90. Given the specifications of the

FRs and the effective threshold, the performance of the CF model can now be evaluated. At each test point, average membership grade for each rule is calculated by plugging the premise parameters of Table B.1. Comparison of the membership grade with the  $\theta$  in the evaluation function of Eq. (3.3) determines the signal. Based on Eq. (B.1), the evaluation function becomes:

$$C_i^* = \begin{cases} 1, & \omega_i^{*1/12} \geq 0.9 \\ 0, & \text{Otherwise} \end{cases} \quad (\text{B.2})$$

If the geometric mean of the membership grades of all the inputs (12 elements) is greater than 0.9, the rule is considered strong and qualifies for decision-making and a decision is made based on the CF model. The weighted average of the regressions for the qualified rules determines the CF output (Eq. (3.8)). Based on the Eq. (3.8), I have:

$$\hat{O}^* = \frac{\sum_{i=1}^{13} \omega_i C_i^* f_i}{\sum_{i=1}^{13} \omega_i C_i^*} \quad (\text{B.3})$$

In the game result forecasting, the match outcome ( $\hat{O}^*$ ) is a weighted average of the regressions specified for each rule (see Table B.2). The weight for each rule's regression is normalized amount of  $\omega_i C_i^*$  as of Eq. (3.12). Finally, for betting purposes the estimated value must be interpreted and classified ( $\hat{O}_{adj}^*$ ) under three labels of “win (+1)”, “draw (0)” or “lose (-1)”. The regression outputs are classified according to the following transfer function:

$$\hat{O}_{adj}^* = \begin{cases} 1, & 0.5 < \hat{O}^* < \infty \\ 0, & -0.5 < \hat{O}^* \leq 0.5 \\ -1, & -\infty < \hat{O}^* \leq -0.5 \end{cases} \quad (\text{B.4})$$

## B.2 IS Performance

The IS accuracy ratios of RVM are presented in Table B.4.

[Table B.4]

### B.3 CF Games

In Table B.5 I present the number of games that the CF has generate a forecast for the relevant exercises.

[Table B.5]

### B.4 ANFIS

ANFIS has been applied successfully in many different aspects of science (Chang and Chang, 2006; Polat and Güneş, 2007; Mathur *et al.*, 2016). ANFIS consists of five layers (see Figure 3.1). Each layer is involved in a specific task of fuzzy modelling through adaptive nodes, pertaining parameters and a processing function, which is updated by a hybrid learning algorithm. In the figure, the adaptive nodes are presented by squares whereas a circle indicates a fixed node. The outcome for each node  $k$  in layer  $j$  is denoted by  $O_k^j$ .

The first layer of ANFIS is made of adaptive nodes where the distance of each input to each rule is estimated through a membership function and reported as membership grade. The membership function as proposed by Jang (1993) for input  $x$  and fuzzy set  $A$  for rule  $i$  is denoted by  $\mu_{A_i}(x)$ :

$$\mu_{A_i}(x) = \exp \left\{ - \left( \frac{x - c_i}{a_i} \right)^2 \right\} \quad (\text{B.5})$$

where  $a_i$  and  $c_i$  are called premise parameters. The most popular membership function is the Gaussian, and this is followed in this study.

The outcome for each node at layer one can be presented as:

$$O_k^1 = \mu_{A_i}(x), \quad k = 1, 2; \quad i = 1, 2 \quad (\text{B.6})$$

$$O_k^1 = \mu_{B_i}(y), \quad k = 3, 4; \quad i = 1, 2 \quad (\text{B.7})$$

In the second layer (nodes  $\Pi$ ), a weight (firing strength) is allocated to each node that represents the power of the associated rule. Therefore, the number of nodes is equal to the number of rules. The weight of rule  $i$  is the product of the

membership grade of the fuzzy set A and the membership grade of the fuzzy set B.

$$O_k^2 = w_i = \mu_{A_i}(x) \times \mu_{B_i}(y), \quad k = i = 1, 2 \quad (\text{B.8})$$

where  $w_i$  is the firing strength of rule  $i$ .

The third layer (nodes N) also involves fixed type nodes with normalising functionality for the firing strength estimated for each fuzzy rule. Thus:

$$O_k^3 = \bar{w}_i = \frac{w_i}{\sum w_i}, \quad k = i = 1, 2 \quad (\text{B.9})$$

The next layer has adaptive nodes. Each node is fed with the inputs and its output is estimated as:

$$O_k^4 = \bar{w}_i f_i(x, y) = \bar{w}_i (p_i x + q_i y + r_i), \quad k = i = 1, 2 \quad (\text{B.10})$$

where  $\bar{w}_i$  is the normalized firing strength for each rule. The parameters set  $\{p_i, q_i, r_i\}$  are called consequent parameters.

The last layer simply aggregates all its inputs. The result is simply the defuzzified ANFIS model realisations:

$$O_1^5 = \sum_{i=1}^2 \bar{w}_i f_i = \frac{\sum_{i=1}^2 w_i f_i}{\sum_{i=1}^2 w_i} \quad (\text{B.11})$$

A hybrid learning algorithm is required to estimate the premise and consequent parameters. The training algorithm of ANFIS is consisted by two stages, a forward and a backwards. In the forward stage, the premise parameters are fixed and the consequent parameters are estimated by the least square method. In the backwards one, the consequent parameters are kept fixed and the errors are backpropagated. Then, the premise parameters are optimised through the gradient descent method (Shapiro, 2002).

## B.5 OP

Consider the case of having  $\mathcal{A}$  potential outcomes for dependent ordinal variable  $y$  under study by a set of independent variables  $x$ . Under a linear assumption, I want to specify the regression model:

$$y_j^* = x_j' \beta + \varepsilon_j \quad (\text{B.12})$$

where for the observation  $j = 1, \dots, T$  a latent continuous variable  $y_j^*$  is estimated by a set of coefficients  $\beta$  and noise term  $\varepsilon_j$ . If I assume that  $\varepsilon_j$  follows a logistic distribution, the model becomes LR type whereas selecting a standard normal distribution derives the OP model. For outcome  $a = 1, \dots, \mathcal{A}$  the connection between the latent variable  $y_j^*$  and the observed ordinal variable  $y_j$  can be presented as:

$$y_j = a \Leftrightarrow \vartheta_{a-1} < y_j^* \leq \vartheta_a \quad (\text{B.13})$$

where  $\vartheta_a$ s are the threshold parameters estimated by training dataset. Let me set the lower bound  $\vartheta_0$  as  $-\infty$ , the upper bound  $\vartheta_{\mathcal{A}}$  as  $+\infty$  and the order as  $\vartheta_0 < \dots < \vartheta_{\mathcal{A}}$ .

In LR one is interested in interpreting the change in the independent variables into the probability of a certain outcome. The conditional probability of ordinal outcome  $a$  is given by:

$$Pr(y_j = a) = Pr(\vartheta_{a-1} < y_j^* \leq \vartheta_a) = Pr(\vartheta_{a-1} < x_j' \beta + \varepsilon_j \leq \vartheta_a) \quad (\text{B.14})$$

By rearranging I have:

$$\begin{aligned} Pr(y_j = a) &= Pr(\vartheta_{a-1} - x_j' \beta < \varepsilon_j \leq \vartheta_a - x_j' \beta) \\ &= Pr(\varepsilon_j \leq \vartheta_a - x_j' \beta) - Pr(\varepsilon_j \leq \vartheta_{a-1} - x_j' \beta) \end{aligned} \quad (\text{B.15})$$

By letting  $\varepsilon_j \sim \mathcal{N}(0,1)$  I obtain:

$$Pr(y_j = a) = Z(\vartheta_a - x_j' \beta) - Z(\vartheta_{a-1} - x_j' \beta) \quad (\text{B.16})$$

where  $Z$  is the standard normal distribution function. The parameters are estimated through a maximum likelihood procedure by a likelihood function  $\mathcal{L}$  as:

$$\mathcal{L}(\vartheta, \boldsymbol{\beta}) = \prod_{j=1}^T \prod_{a=1}^{\mathcal{A}} [Z(\vartheta_a - x'_j \boldsymbol{\beta}) - Z(\vartheta_{a-1} - x'_j \boldsymbol{\beta})]^{z_{j,a}} \quad (\text{B.17})$$

where  $z_{j,a}$  is an indicator variable equals to one when the ordinal outcome  $a$  for the sample  $j$  is observed ( $y_j = a$ ) or zero otherwise. The maximum likelihood estimates are given then by:

$$\frac{\partial \ln \mathcal{L}}{\partial \vartheta} = 0 \text{ and } \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = 0. \quad (\text{B.18})$$

The estimated conditional probability for each outcome ( $\hat{\pi}_{j,a}$ ) is given by plugging in the optimal parameters ( $\vartheta^*$  and  $\boldsymbol{\beta}^*$ ) from Eq. (B.18).

$$\hat{\pi}_{j,a} = Z(\vartheta_a^* - x'_j \boldsymbol{\beta}^*) - Z(\vartheta_{a-1} - x'_j \boldsymbol{\beta}^*) \quad (\text{B.19})$$

Finally, for each observation  $j$ , predicted outcome is the one with maximum  $\hat{\pi}_{j,a}$  as:

$$\hat{y}_j = \arg \max_a (\pi_{j,a}). \quad (\text{B.20})$$



## Appendix C

### C.1 Monte Carlo simulations

In this Appendix, I present supporting evidence of the finite sample performance of the  $\text{DFDR}^{+/-}$  test using a Monte Carlo experiment. The main goal is the exploration of the empirical level and power of the test in accurately estimating the proportions of outperforming, underperforming and neutral trading rules. Even though I am mainly focused on the FDR rate and its power on the rejection frequency of rules with significant returns (either positive or negative), I also compare it with the power of the FWER rate and especially with the RW method.

Before setting up the simulations it is necessary to ensure that my experiment correctly embodies the empirical properties of the technical trading strategies employed, such as their time series and cross-sectional dependencies (see also Barras *et al.*, 2010; Hsu *et al.*, 2010; Bajgrowicz and Scaillet, 2012). I have previously demonstrated that the technical trading rules are fully characterized by a weak form of dependence, this holds especially for those belonging in the same family (e.g. MAs). This is the main property I need to take that into consideration when constructing my experiment. By this way, I can also examine whether the  $\text{DFDR}^{+/-}$  has indeed a good response to weak dependence conditions. Towards this direction and I resample simultaneously matrices of  $\ell \times l$  returns, where  $\ell$  the random block size of is consecutive time series observations ( $\bar{\ell} = 10$ ) under the stationary bootstrap and  $l = 21,195$  denotes the trading rules universe as in the empirical exercise. This approach also allows me to preserve the cross-sectional dependencies among the strategies of the same class, while I preserve also autocorrelation every time the same bootstrap replication is applied to all trading rules. For the Monte Carlo experiment, I randomly select the original 155-day sample (i.e. seven months) from 1 July 2013 to 1 February 2014 to simulate the trajectories and I generate the 155-day trajectories for the  $l = 21,195$  trading rules as in the empirical exercise. I employ the stationary bootstrap to create every realized trajectory similar to calculating the  $p$ -values of the empirical study. I generate 1,000 bootstrap replications of returns, where each replication has similar statistical properties.

In order to obtain the true power of  $DFDR^{+/-}$  test in selecting the proportions of outperforming, underperforming and neutral rules I need to control these proportions beforehand, likewise observing them a priori. I can then compare them with their corresponding estimations based on the  $DFDR^{+/-}$ . I adjust 20% of the simulated strategies to outperform the benchmark, 50% to deliver “neutral” returns with no significant performance and 30% to underperform the benchmark during the simulation process. The selected outperforming (underperforming) strategies consist only of neighbouring rules, ranked in terms of highest (lowest) returns in the sample to avoid ending up with a group of similar outperforming or underperforming having slightly different parameters.

In terms of the specific procedure followed, I achieve the control of outperforming, “neutral” and underperforming rules by recentering the generated returns of each trading rule with its own mean and I utilize that across all five families of rules. This leads to all trajectories having almost zero-mean properties but retaining their corresponding, unique standard deviations. I then shift the paths of the outperforming and underperforming rules by some positive and negative value respectively, while keeping each rule’s corresponding standard deviation the same. Such a parallel transition does not affect empirical properties of the paths, other than the mean (see Paparoditis and Politis, 2003). The notion is to carry the trajectories of different strategies in such a way as to exactly acquire the same, positive Sharpe ratio for all outperforming rules and the same negative Sharpe ratio for all underperforming rules. I specify both chosen positive and negative Sharpe ratios in advance<sup>43</sup>.

As about the target Sharpe ratios employed for shifting the paths of outperforming and underperforming strategies, I follow the study of Bajgrowicz and Scaillet (2012) and select Sharpe ratios closely related to the ones obtained in this Chapter’s empirical exercise. I set three specific targets of positive Sharpe ratios for the outperforming rules, i.e., 2, 3, 4; and three specific targets of negative Sharpe ratios for the underperforming rules, i.e., -2, -3, -4. All of them correspond to annualized Sharpe ratios as those calculated from the daily returns

---

<sup>43</sup> I multiply the corresponding standard deviation of each rule by the pre-specified Sharpe ratio and I add up the calculated value to each data point so that the mean for the rule becomes Sharpe ratio times sigma.

of each strategy. I then consider pairs of the Sharpe ratios above in order to adjust the outperforming and underperforming rules, while shifting their trajectories towards the target. Take the (2, -2) pair, for example, I design 20% of the rules to yield an equal Sharpe ratio of 2 (i.e., outperforming) and likewise, all 30% of the rules share an equal Sharpe ratio of -2 (i.e., underperforming). The rest 50% of this Chapter's trading universe show zero performance. This results in nine possible combinations of positive and negative Sharpe ratio pairs representing fixed alternative hypotheses against the null of a Sharpe ratio being equal to zero. The above levels seem to match 4<sup>th</sup> Chapter's historical sample results since I obtain positive annualized Sharpe ratios up to 4 for the best-performing strategies and negative annualized Sharpe ratios down to -4 for the worst-performing ones. However, the outperformance versus underperformance pair of (2, -2) still portrays a quite challenging setup for my portfolio construction method.

I present the results of this Chapter's Monte Carlo experiments in Tables C.1 to C.3 below. Table C.1 displays the annualized mean excess return quartiles for the controlled outperforming and underperforming technical trading rules based on the 1,000 Monte Carlo replications for the nine combinations of Sharpe ratios examined.

### [Table C.1]

In general, the annualized mean returns I obtain seem quite analogous to their corresponding Sharpe ratio levels, either positive or negative.

Focusing on the estimation power of DFDR<sup>+/-</sup> approach, Table C.2 presents the estimates for the proportions of outperforming ( $\widehat{\pi}_A^+$ ), underperforming ( $\widehat{\pi}_A^-$ ) and neutral ( $\widehat{\pi}_0$ ) strategies under the Sharpe ratio metric and for the nine possible Sharpe ratio pairs. It also reports the success of the estimators in tracking the actual proportions of outperforming ( $\pi_A^+ = 20\%$ ), underperforming ( $\pi_A^- = 30\%$ ), and neutral ( $\pi_0 = 50\%$ ) trading rules. I employ the "point estimates method" of Storey et al, (2004) to obtain the estimators of these proportions based on the Monte Carlo results. Based on a series of Monet Carlo experiments using the DFDR<sup>+/-</sup> test in this Chapter's technical trading rules universe, this time I keep the

cut-off threshold fixed to  $\gamma^* = 0.4$ , as at this point  $\widehat{\pi}_A^+$  and  $\widehat{\pi}_A^-$  become constant. In other words, they include both genuine and false selections of trading rules and so represent the total number of outperforming and underperforming rules respectively.

[Table C.2]

The DFDR<sup>+/-</sup> approach seems to provide quite robust estimators for the outperforming, underperforming and neutral proportions of technical trading rules, with only small deviations from their true corresponding ones. For instance, looking the (3, -3) Sharpe ratios pair, the estimator for the outperforming rules (i.e.,  $\widehat{\pi}_A^+$ ), is 15.23%, the relevant estimator for underperforming rules (i.e.,  $\widehat{\pi}_A^-$ ) is 27.68% and the one for neutral rules (i.e.,  $\widehat{\pi}_0$ ) is 57.09%, which are quite close to their true levels of 20%, 30% and 50% respectively. This clearly highlights the power of this Chapter's method in accurately identifying the true proportions of outperforming, underperforming and neutral rules in the entire population.

Finally, I present in Table C.3 the performance of constructed portfolios of outperforming rules under the DFDR<sup>+</sup> approach based on the Monte Carlo simulation and for each of the nine Sharpe ratio combinations. I control the DFDR<sup>+</sup> at a prespecified level similar to this Chapter's empirical exercise. For instance, I build two different types of DFDR<sup>+</sup> portfolios by setting the targets of erroneous selections at 10% and 20% respectively. In terms of performance and power, the table reports the actual FDR achieved (FDR<sup>+</sup>) in comparison with its fixed level adjusted in advance (i.e., 10% and 20%), the proportions of genuinely best-performing rules over the total number of outperforming rules denoted as "power", and the absolute number of genuinely best-performing trading rules as "portfolio size"<sup>44</sup>. As mentioned before and for comparison purposes against an FWER method, I also run the same experiment and findings using the RW test, while controlling the FWER at the 5% and 20% level respectively.

---

<sup>44</sup> I compute the actual FDR for the positive tail (FDR<sup>+</sup>) by replacing the actual proportion of neutral trading rules (i.e.,  $\pi_0 = 50\%$ ) instead of the estimated one (i.e.,  $\widehat{\pi}_0$ ) in  $FDR^+ = \frac{\pi_0 \times l \times \gamma / 2}{\#\{p_k \leq \gamma, \varphi_k > 0; k=1, \dots, l\}}$

[Table C.3]

The findings of Table C.3 reveal that the DFDR<sup>+</sup> approach is superior in terms of finite sample power than the more conservative FWER approaches such as the RW approach. Specifically, the DFDR<sup>+</sup> reports a robust power in rules selection and portfolio size, while it closely tracks the actual false discovery rate across all conditions and Sharpe ratio pairs. For example, consider again the (3, -3) Sharpe ratios pair, the 10%-DFDR<sup>+</sup> portfolio efficiently converges to its FDR rate at 8% and successfully discovers on average 64.74% of the best-performing rules. On the other hand, the relevant 5%-RW portfolio discovers only 0.01% of the best-performing rules on average, while it meets its target rate only at 0.04%. When it comes to portfolios' sizes the 10%-DFDR<sup>+</sup> outstandingly outperforms the 5%-RW approach by sufficiently selecting 3,048 rules, while the FWER method detects only 0.6. Increasing the target rate of the FWER to 20% doesn't improve this picture since it slightly improves the power of selection to 0.07% and the portfolio size to 3.47 rules. The 20%-DFDR<sup>+</sup> though performs even better by detecting on average 66.3% of the outperforming rules and forms a portfolio of 3,321 trading rules. In terms of target rate, the 20%-DFDR<sup>+</sup> portfolio falls below 20% and achieves an FDR<sup>+</sup> of 10.59%. Asymptotic theory is the most possible reason for this outcome, but the 20%-DFDR<sup>+</sup> portfolio is still able to successfully deal with data snooping bias as seen above. In overall, this Chapter's Monte Carlo experiments undoubtedly reveal that the DFDR<sup>+/−</sup> method has a greater power over conservative FWER methods, such as the RW procedure, especially in a big-data framework, while they highlight the main drawback of the FWER methods, such as the RW approach, which terminate as soon as a false rejection is discovered, even in cases I allow for a bigger target rate (i.e. 20%).

## C.2 Robustness Exercise

The exercises in Sections 4.4 and 4.5 are repeated by setting the IS period to 1 year and the OOS to 3, 6 and 9 months, respectively. Figure C.1 and Tables C.4 and C.5 address the IS studies while Tables C.6 to C.8 report corresponding results for the OOS.

Tables C.4 and C.5 present the same analysis as in Section 4.4.1 for the IS of one year. Table C.4 presents the percentage and standard deviations of the

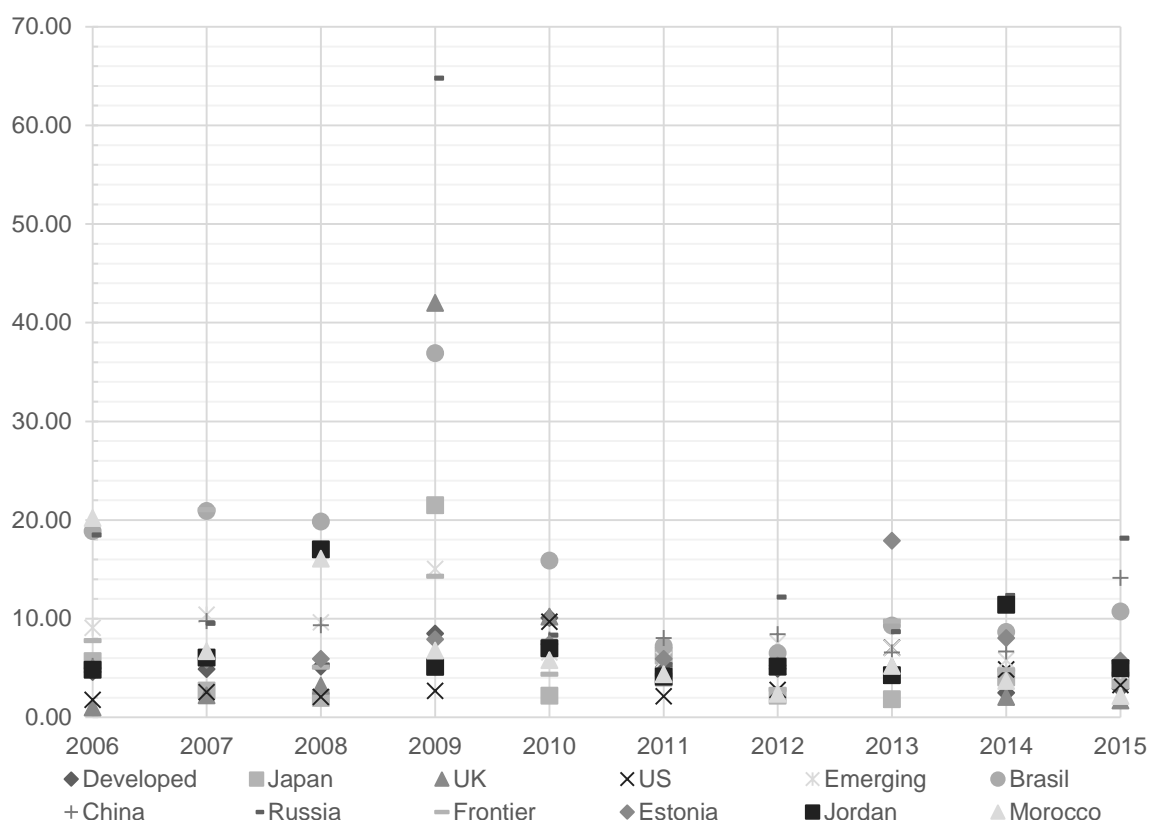
survivor rules identified by the 10%-DFDR<sup>+</sup> approach, analogous to Table 4.2. Likewise, Table C.5 displays the annualized returns and Sharpe ratios of significant rules after one-way transaction costs, analogous to Table 4.3.

### [Tables C.4 and C.5]

Comparing the corresponding Tables 4.2 and 4.3 to Tables C.4 and C.5, I conclude that using one year as this Chapter's backtesting period the performance of trading rules is quite better in terms both of excess mean return and Sharpe ratio criteria. This can be expected since the bigger the sample period the more exposed are the technical trading rules to fluctuations and market risks in general, leading to lower performance most of the times. Realizing higher IS returns in smaller sample horizons is a common phenomenon not only for the relevant literature but also for the trading desks.

As a robustness check for Section 4.4.2, Figure C.1 presents the break-even transaction cost for an IS period of one year, analogous to Figure 4.1. The average break-even cost for the two-year and one-year cases are 9% and 8% respectively. The highest performance in Figure C.1 belongs to Russia with 65% in 2009. This is comparable to the two-year case where the best performance – 51% – is reported in 2009 by Japan. The lowest performance in Figures 4.1 and C.1 is almost the same and equal to 1%. Thus, comparison of the results from the two figures show that the average level of break-even cost is higher for the two-year IS, but the performance range is wider for the one-year case. The robustness check for the transaction cost shows that longer IS period leads to detection of more stable pattern even though short-term approach might lead to a higher temporary reward.

**Figure C.1. Break-even Cost for the Top Performing Survivor of the DFDR+ Procedure (IS 1 Year)**



**Note:** The values are in percentages and calculated as the transaction cost that sets the excess return to zero over the period under study. The IS period is set to one year. The values are calculated by repeating the procedure at the start of each month and averaging over 12 months.

From technical trader point of view, what matters most is the OOS simulation findings rather than the IS ones. Tables C.6 to C.8 correspond to one year as an IS period while considering the OOS periods of one, three and six months respectively.

### [Table C.6 to C.8]

In this case, I observe opposing evidence with regards to the IS period chosen each time. For example, comparing Tables 4.4 to 4.6 corresponding to the IS period of two years and the same OOS periods with Tables C.6 to C.8, the results are in favour with the first approach. I may attribute these findings to including more information (larger historical sample) when searching for a predictive technical trading rule IS resulting in better OOS performance. Moreover, I can also justify the above findings to technical trading rules' specific characteristics and parameterizations. For instance, I utilize technical trading rules, even of the same

family, whose lagged values span from one day up to one year (e.g. a double MA of two and five days respectively; a double MA of 150 and 250 days respectively). This means that they need different learning times in order to capture all the available market trends, momentum or reversals. Choosing a small IS period (i.e. one year) might provide enough information for trading rules utilizing short periods of previous market returns (i.e. a double MA of two and five days) but not enough inputs for trading rules looking back at longer periods of market movements (i.e. a double MA of 150 and 250 days respectively). Hence, in my opinion considering a sufficient enough horizon based on a strategy's properties, while setting an optimal IS, OOS ratio, is equally important to the selection of the best predictive rule.

I investigate further the above optimality in the IS and OOS ratio with respect to sample periods chosen by looking the corresponding performances of the significant technical trading rules over the three different OOS periods (i.e. one, three and six months) examined and the IS period of one year in this Appendix. In terms of average annual performance of all markets considered (i.e. last row), Tables C.6 to C.8 reveal specific patterns in OOS excess profitability of technical trading rules according to both mean return and Sharpe ratio metrics. Specifically, from 2006 to 2009 employing the short OOS period of one month achieves higher mean returns as well as Sharpe ratios compared to the longer periods used (i.e., three and six months), which display a decay as the OOS periods becomes larger during these years. On the contrary, there is a turning point in this phenomenon for the 2010-2012 period. The longer the OOS period the better greater the mean return and Sharpe ratio. Despite that, I must note that both metrics appear negative during these years. For the rest of the years (i.e., 2013-2015), technical trading rules seem to perform better using the OOS period of one month, yielding even positive metrics in 2015. In general, profitability diminishes as I approach the most recent periods for all OOS periods and across all markets considered. This evidence is consistent with the general findings presented in Section 4.5.1 when a two-year IS horizon was considered.

When it comes to each market's average performance over the full ten-year period (i.e., last column) the picture is quite different. There is no clear evidence towards the support of a specific OOS period in general and sometimes both performance metrics employed provide contradictory results. I conclude that the



most suitable OOS horizon depends on the specific market studied. For developed markets, the performance of trading rules seems to improve according to the Sharpe ratio as I expand the OOS period, but this is not the case when the mean excess return is adopted as the performance criterion. As about the emerging markets and frontier markets results provide an opposing order, in which the shorter OOS periods employed the better technical trading rules performance, which is consistent with both mean return and Sharpe ratio. Despite that, I must also mention that technical trading rules underperform the benchmark most of the times, especially in the developed markets, which once again justifies the usage of the IS period of two years.

## Appendix D

This Appendix includes characteristics of the volatility models used, the fine steps taken in the methodology section and remainder results for the five loss functions not presented in the main text. Initially, the volatility models' pool is specified in Appendix D.1. The pseudo-algorithm of the DFDR<sup>+</sup> approach is provided in Appendix D.2. The models' performance density distributions is given in Appendix D.3. Appendix D.4 presents the true discoveries dynamics over time for different distributions and the error distribution analysis results are shown (Appendix D.5). The mean and conditional variance robustness checks are presented in Appendices D.6 and D.7 respective. Finally, Appendix D.8 presents the results for different classes studied.

### D.1 Specification of the Pool

Table D.1 provides the characteristics of the models under study. I study twenty classes of volatility models from four families (GARCH, SV, EWMA, and HAR), up to four distributions<sup>45</sup> for innovation distribution fitted for the standardized return process, three specifications for mean and two definitions of the conditional variance.

[Table D.1]

### D.2 Model Selection Algorithm

The following procedure is followed to find the superior models compared to the benchmark. The steps are the same for all 6 loss functions.

1. Calculate the centred test statistics given the loss functions in Eq.s (5.22 to 5.27) and the benchmarks (ARCH (1), GARCH (1,1), or 90<sup>th</sup> percentile of the pool).
2. Generate  $B = 1000$  stationary bootstrap to generate the set of centered null test statistics  $\varphi'_{b,i}$ , for  $b = 1, \dots, B$ .
3. Calculate the p-values ( $\hat{p}_i$ ) for  $m = 1512$  models.

---

<sup>45</sup> The combination of GED and SV generated poor fitting for common optimization algorithms. Therefore, I discarded this combination from the pool to meet my computational constraints.

4. Estimate the optimal tuning parameter  $\lambda^*$  and estimated null proportion

$\hat{\pi}_0(\lambda^*)$  as:

- a. Select the number of  $\lambda$  support points ( $n = 20$ ).
- b. Define the  $\Lambda = \{\lambda_1, \dots, \lambda_n\}$
- c. Compute the estimated null proportion as  $\hat{\pi}_0(\lambda_l) = \frac{\#\{\hat{p}_i > \lambda_l\}}{(1-\lambda_l)m}$  for  $i = 1, \dots, m$  and  $l = 1, \dots, n$
- d. Find the first  $l$  where  $\hat{\pi}_0(\lambda_l) \geq \hat{\pi}_0(\lambda_{l-1})$

5. Choose a target FDR<sup>+</sup> controlling level ( $\alpha$ ) e.g. 10%

6. Sort the p-values in an ascending order where  $\hat{p}_{R_1} \leq \dots \leq \hat{p}_{R_m}$

7. Define step by  $j = 1$  and corresponding significance region with  $\gamma'_j = \hat{p}_{R_j}$

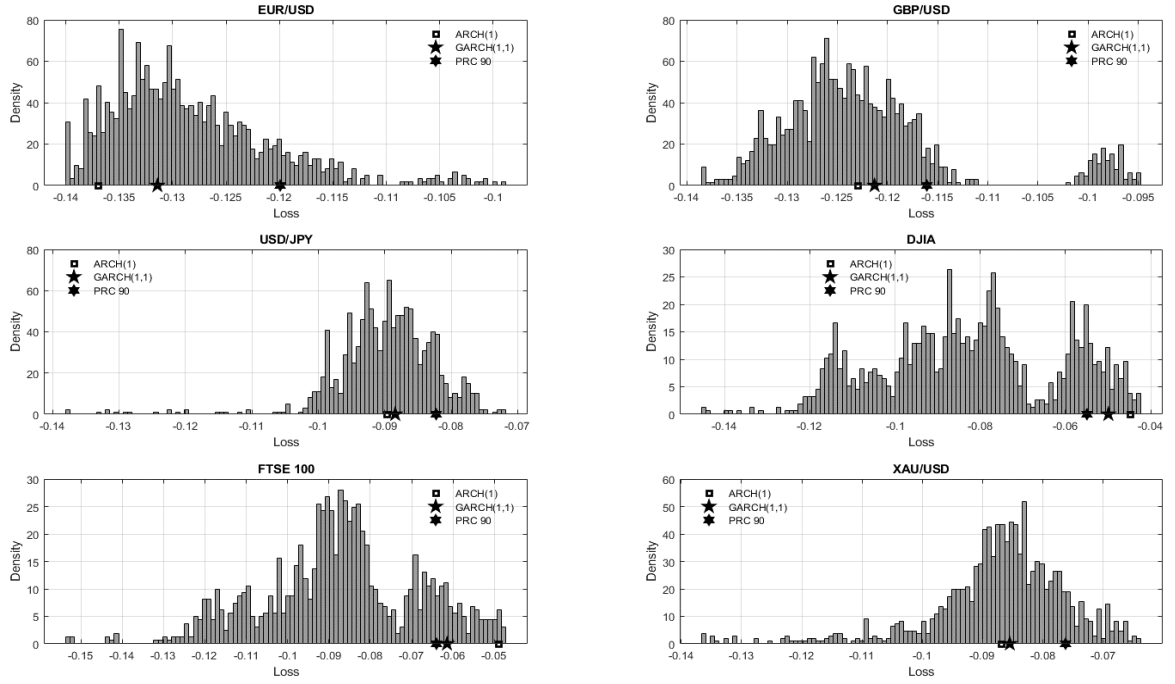
8. Compute the  $\widehat{FDR}_{\gamma'_j}^+(\hat{\pi}_0)$ .

- a. If the  $\widehat{FDR}_{\gamma'_j}^+ < \alpha$ 
  - i. Reject all  $P_i \leq P_{R_j}$
  - ii. Set  $j = j + 1$ ,  $\gamma'_j = \hat{p}_{R_j}$  and go back to 8.
- b. Otherwise if the  $\widehat{FDR}_{\gamma'_j}^+ \geq \alpha$ 
  - i. Reject all  $\hat{p}_i \leq \hat{p}_{R_j}$
  - ii. Terminate the process

### D.3 Densities for Other loss Functions

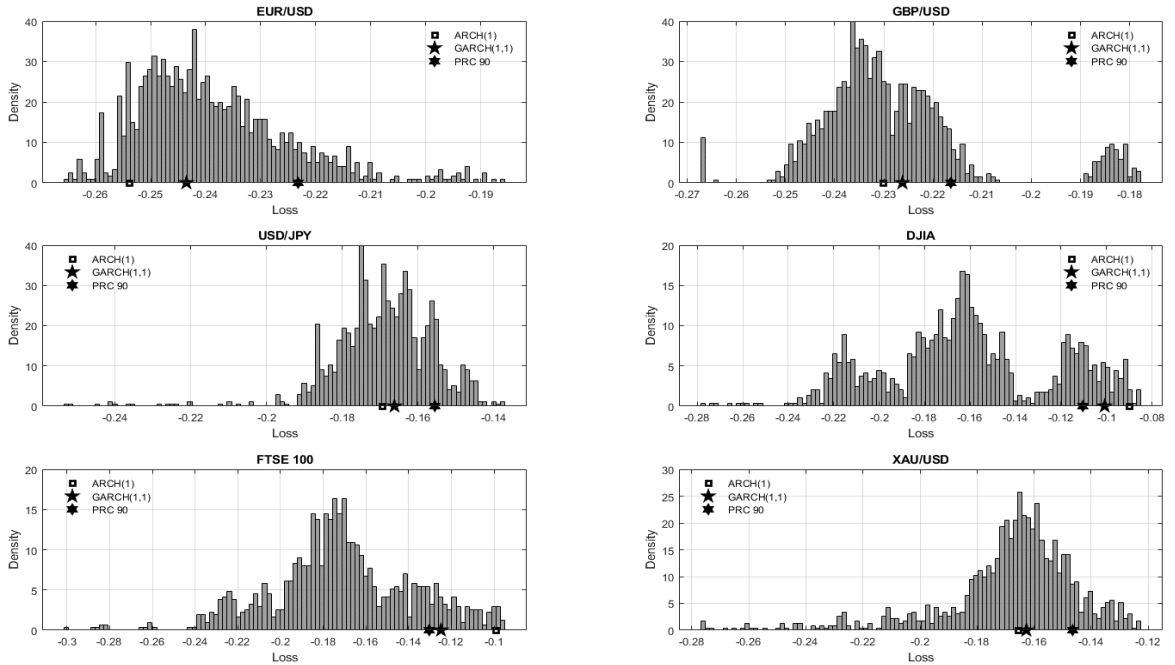
This Appendix presents the average performance of the volatility pool over the whole study period (2013-2017). The performance is measured based on six loss functions. The  $MSE_1$  is presented in the main text. The other five loss functions provided here are  $MAE_1$ ,  $MAE_2$ ,  $MSE_2$ ,  $R^2\text{LOG}$ , and  $QLIKE$ . The specifications of the loss functions are provided in Table 5.3.

Figure D.1: Model Performance Density for  $MAE_1$



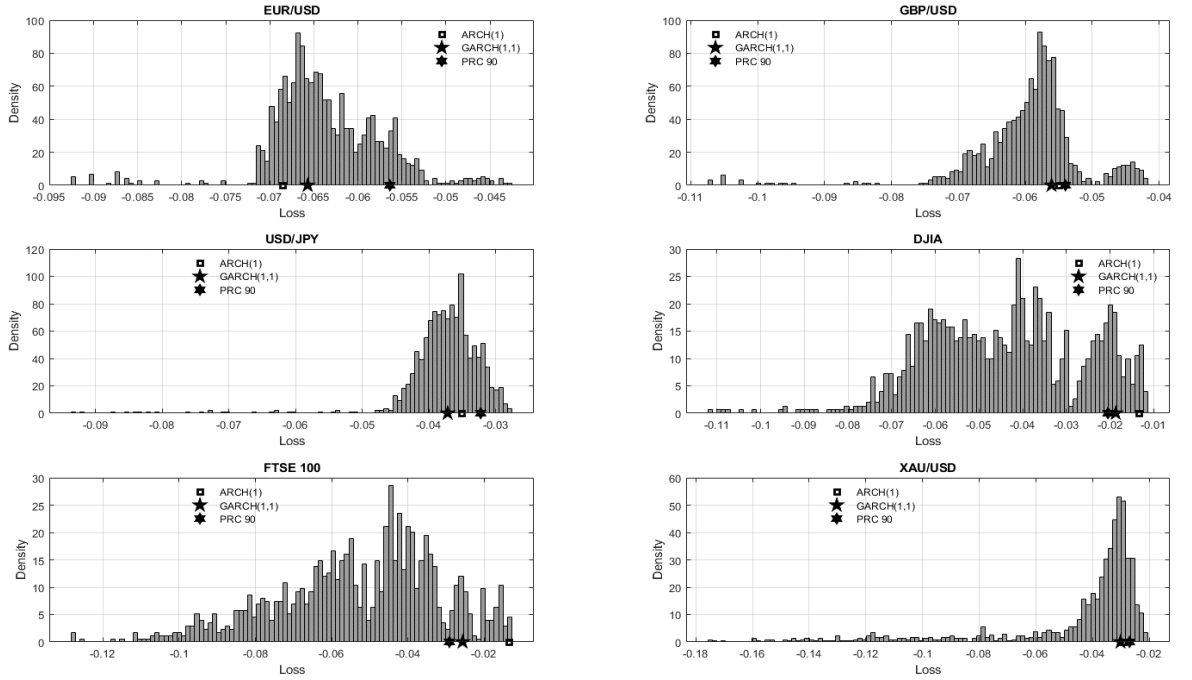
**Note:** The figures present the density of average loss across the pool based on the  $MAE_1$  benchmark. The  $x$  axis is  $\bar{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  is the OOS subperiods of one year. The polygonal present the location of the benchmarks studied. The square, pentagon, and hexagon correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

Figure D.2: Model Performance Density for  $MAE_2$



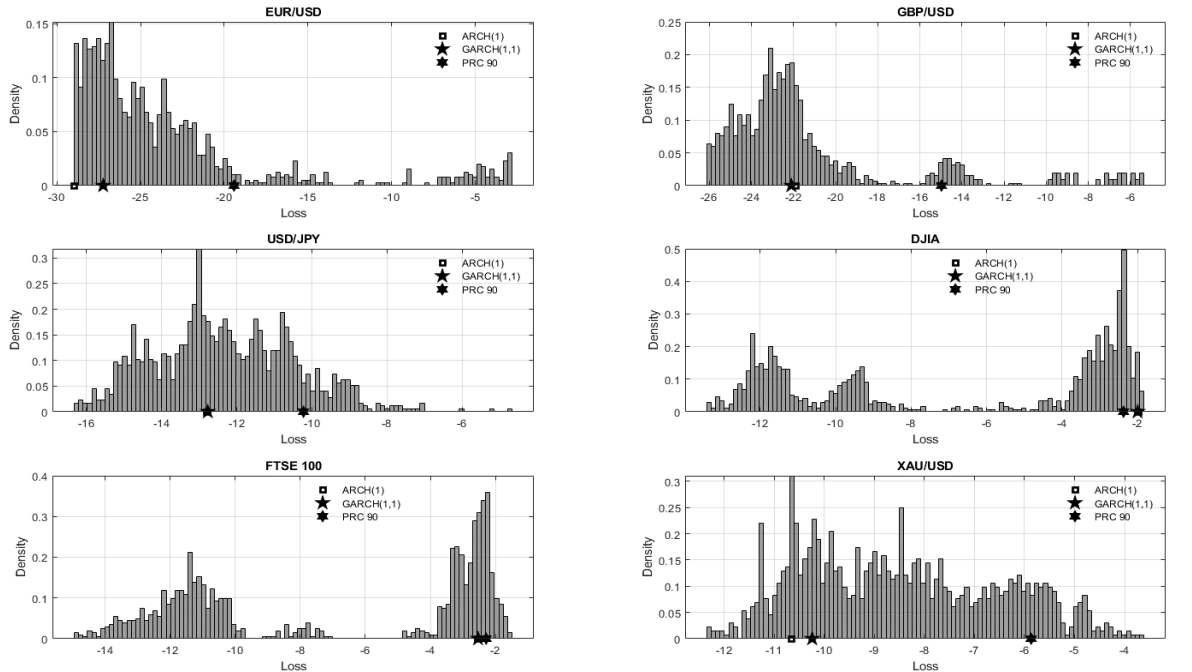
**Note:** The figures present the density of average loss across the pool based on the  $MAE_2$  benchmark. The  $x$  axis is  $\bar{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  is the OOS subperiods of one year. The polygonal present the location of the benchmarks studied. The square, pentagon, and hexagon correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

Figure D.3: Model Performance Density for  $MSE_2$



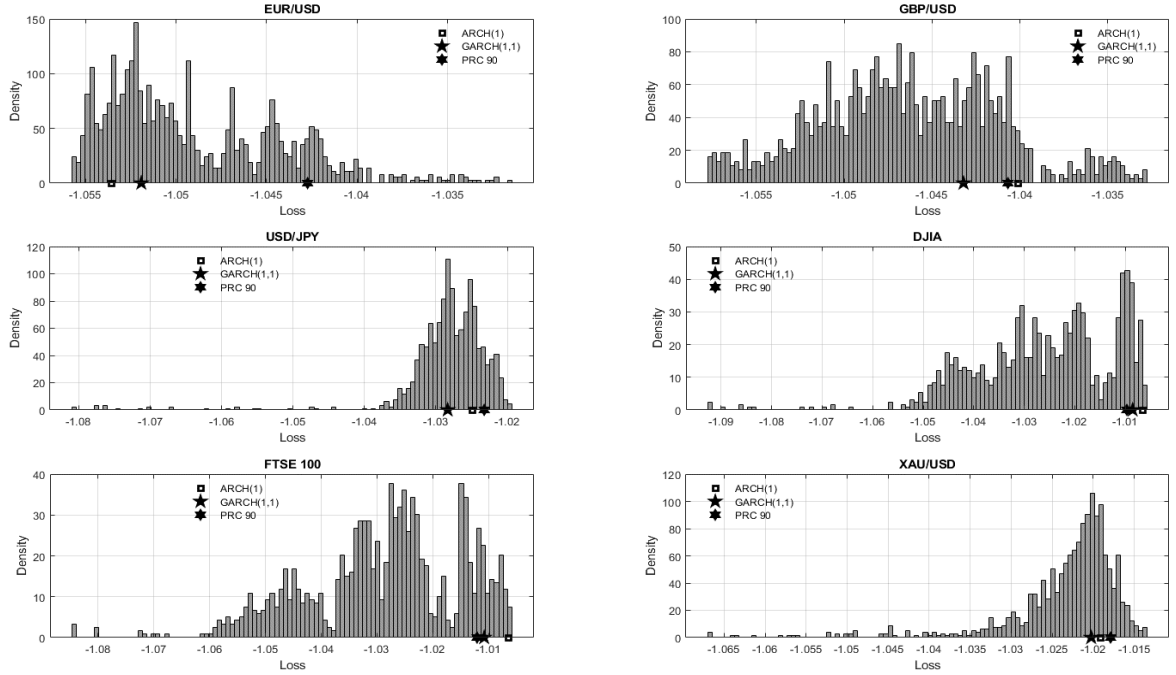
**Note:** The figures present the density of average loss across the pool based on the  $MSE_2$  benchmark. The  $x$  axis is  $\tilde{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  is the OOS subperiods of one year. The polygonal present the location of the benchmarks studied. The square, pentagram, and hexagram correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

Figure D.4: Model Performance Density for  $R^2LOG$



**Note:** The figures present the density of average loss across the pool based on the  $R^2LOG$  benchmark. The  $x$  axis is  $\tilde{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  is the OOS subperiods of one year. The polygonal present the location of the benchmarks studied. The square, pentagram, and hexagram correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

Figure D.5: Model Performance Density for QLIKE



**Note:** The figures present the density of average loss across the pool based on the QLIKE benchmark. The  $x$  axis is  $\tilde{\varphi}_i = 1/5 \sum_{s=1}^5 (-\mathcal{L}_{i,s})$  where  $s$  is the OOS subperiods of one year. The polygonal present the location of the benchmarks studied. The square, pentagram, and hexagram correspond to the ARCH (1), GARCH (1,1), and PRC 90 respectively.

## D.4 True Discoveries Dynamics over Time for Other Distributions

In this Appendix, the number of the true discoveries for each year over 2013-2017 are presented. The true discoveries are the rejections of the DFDR<sup>+</sup> tests based on six loss functions. The results for the MSE<sub>1</sub> are presented in the main text. The other five loss functions provided here are MAE<sub>1</sub>, MAE<sub>2</sub>, MSE<sub>2</sub>, R<sup>2</sup>LOG, and QLIKE. The specifications of the loss functions are provided in Table 5.3.

[Tables D.2 to D.6]

## D.5 Error Distribution Analysis for Other Loss Functions

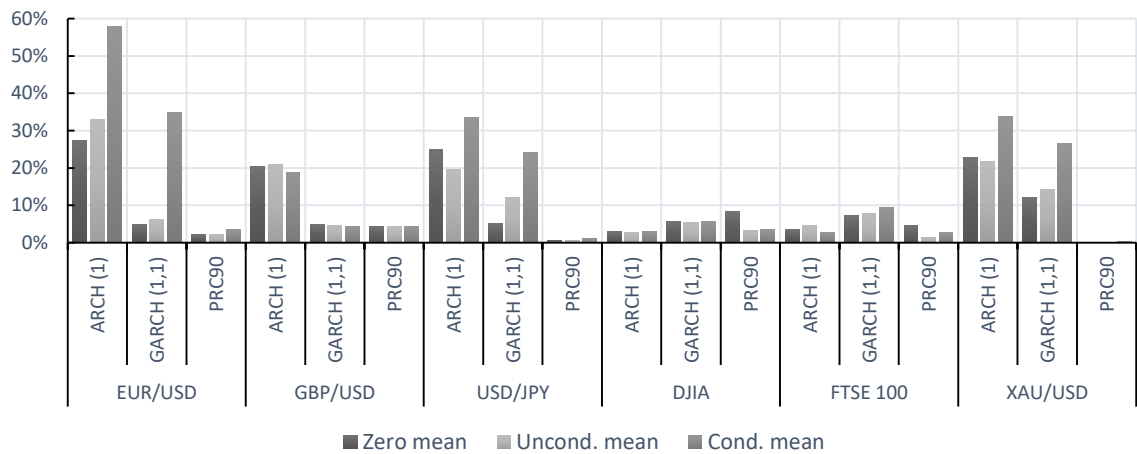
This Appendix focuses on the fitted innovations distribution used to generate the standardized return series. The Tables D.7 to D.11 present the proportion of the models with each error distribution that survive the tests. The values correspond to averages over the whole study period (2013-2017). The test survivors are based on six loss functions. The results for the MSE<sub>1</sub> are presented in the main text. The other five loss functions provided here are MAE<sub>1</sub>, MAE<sub>2</sub>, MSE<sub>2</sub>, R<sup>2</sup>LOG, and QLIKE. The specifications of the loss functions are provided in Table 5.3.

[Tables D.7 to D.11]

## D.6 Mean Analysis of Other Loss Functions

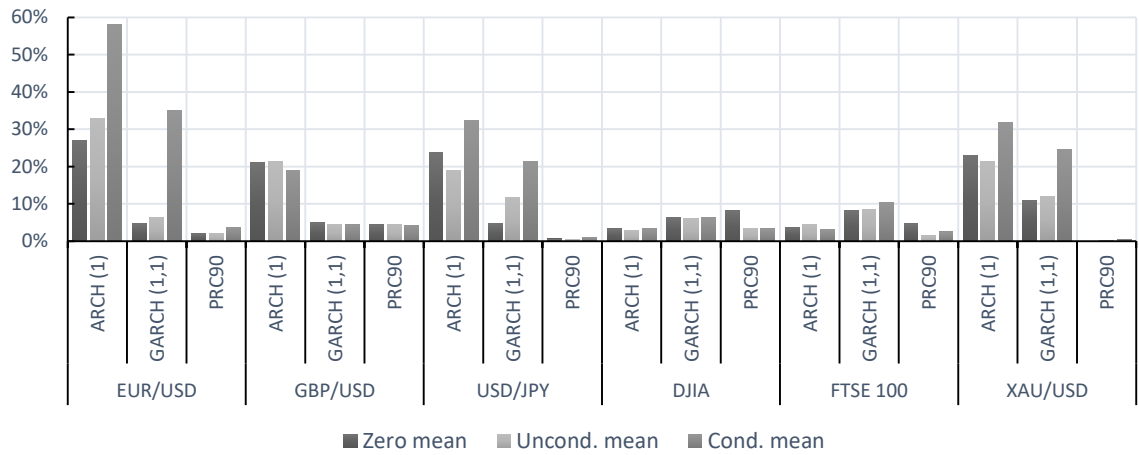
This Appendix studies the success rate for three specifications of the conditional mean used to generate the standardized return series. The Figures D.6 to D.10 present the proportion of the models with each estimated mean specification that survive the tests. The values correspond to averages over the whole study period (2013-2017). The test survivors are based on six loss functions. The results for the  $MSE_1$  are presented in the main text. The other five loss functions provided here are  $MAE_1$ ,  $MAE_2$ ,  $MSE_2$ ,  $R^2LOG$ , and  $QLIKE$ . The specifications of the loss functions are provided in Table 5.3.

**Figure D.6: Conditional Mean Survival Proportion Dynamics Across the Markets for  $MAE_1$**



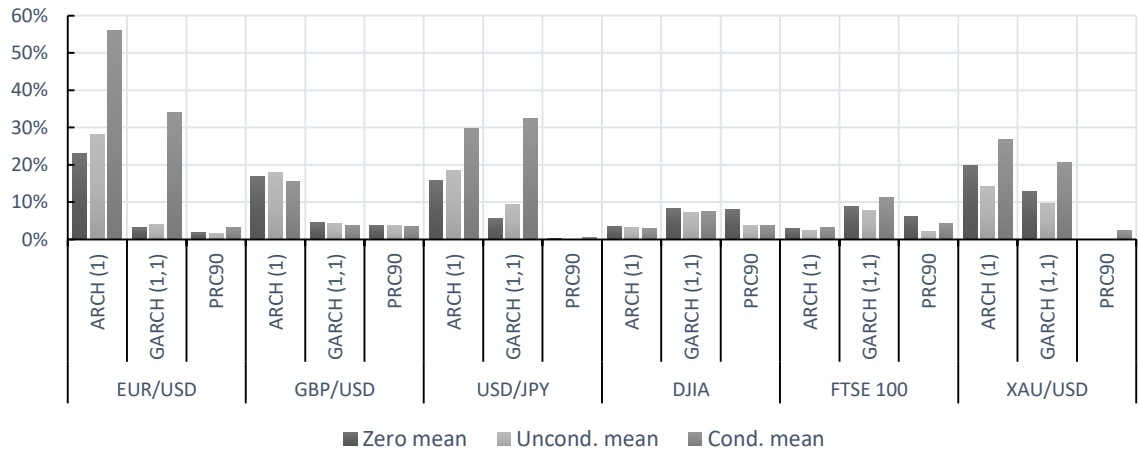
**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MAE_1$ .

Figure D.7: Conditional Mean Survival Proportion Dynamics Across the Markets for  $MAE_2$



**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MAE_2$ .

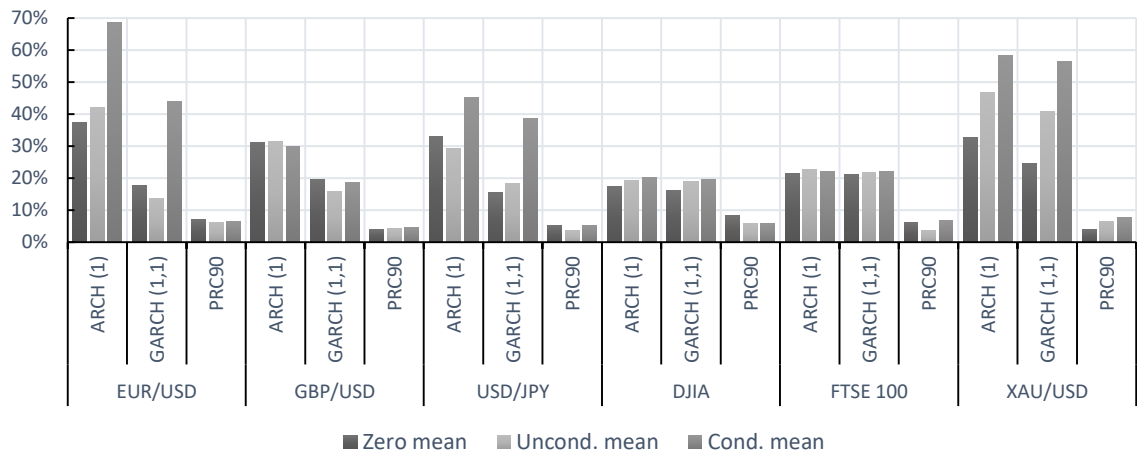
Figure D.8: Conditional Mean Survival Proportion Dynamics Across the Markets for  $MSE_2$



**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_2$ .

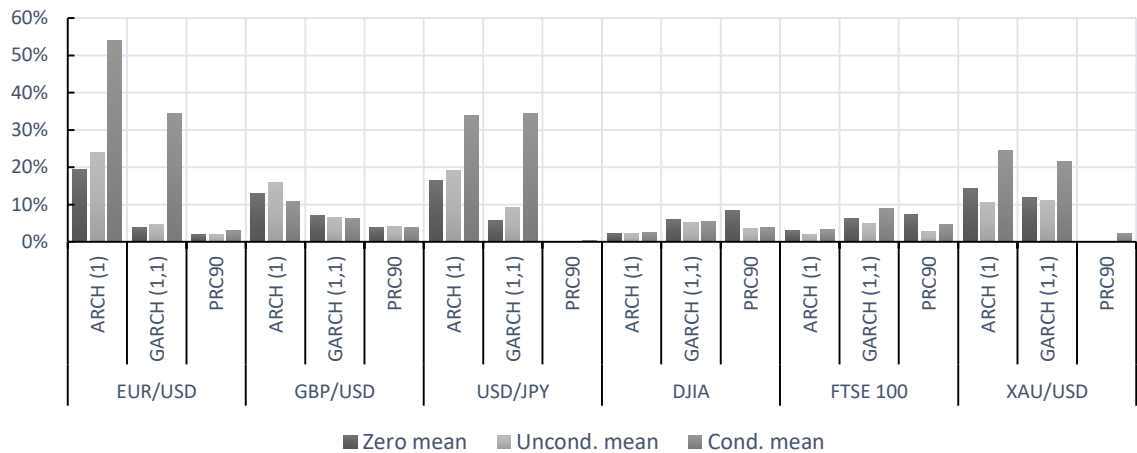


**Figure D.9: Conditional Mean Survival Proportion Dynamics Across the Markets for  $R^2\text{LOG}$**



**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $R^2\text{LOG}$ .

**Figure D.10: Conditional Mean Survival Proportion Dynamics Across the Markets for QLIKE**



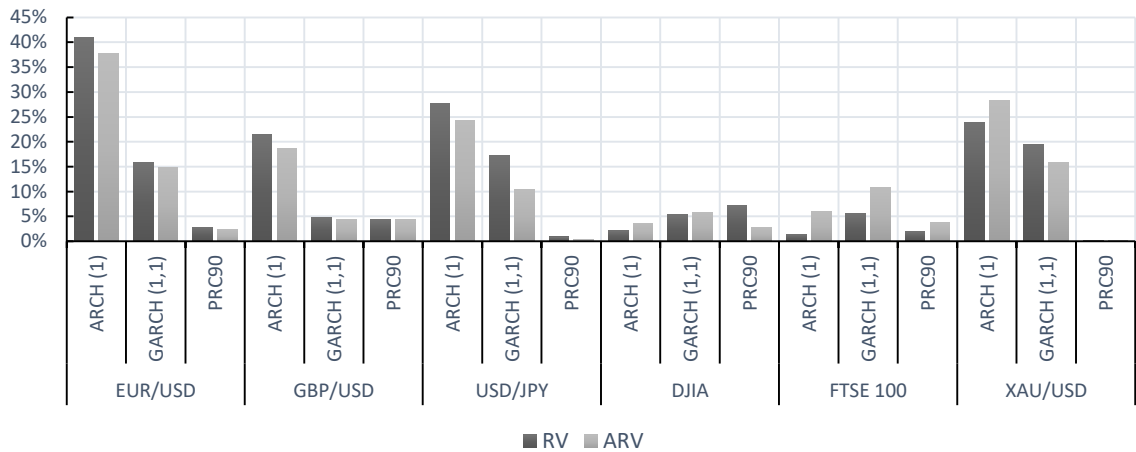
**Note:** The figure presents the average proportion of models with each mean estimation choice. The 'Uncond.' and 'Cond.' stand for unconditional and conditional mean specifications. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is QLIKE.

## D.7 Conditional Variance Analysis for Other Loss Functions

This Appendix examines the role of the conditional variance used to generate the standardized return series. The Figures D.11 to D.15 present the proportion of the models with either of two variance approximations that survive the tests. The values correspond to averages over the whole study period (2013-2017). The test

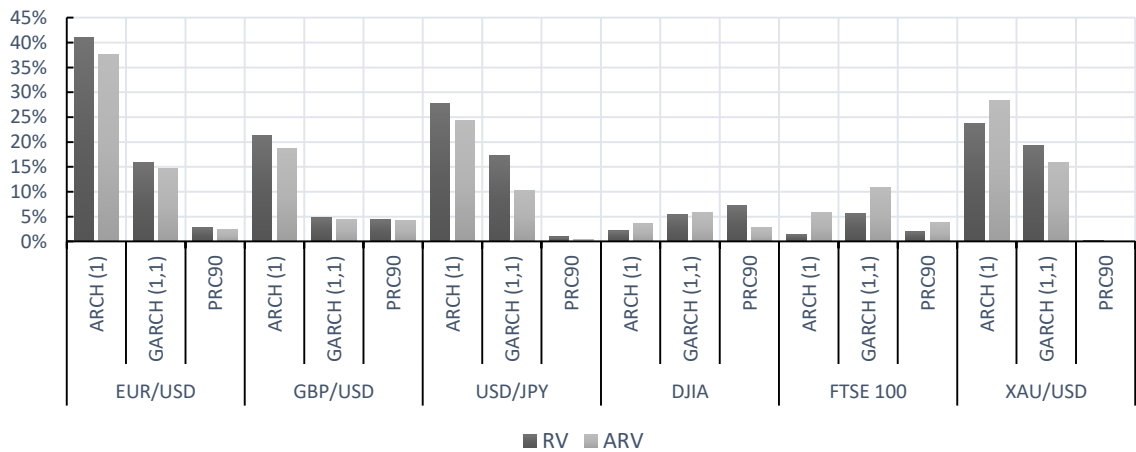
survivors are based on six loss functions. The results for the  $MSE_1$  are presented in the main text. The other five loss functions provided here are  $MAE_1$ ,  $MAE_2$ ,  $MSE_2$ ,  $R^2LOG$ , and  $QLIKE$ . The specifications of the loss functions are provided in Table 5.3.

**Figure D.11: Conditional Variance Survival Proportion Dynamics Across the Markets for  $MAE_1$**



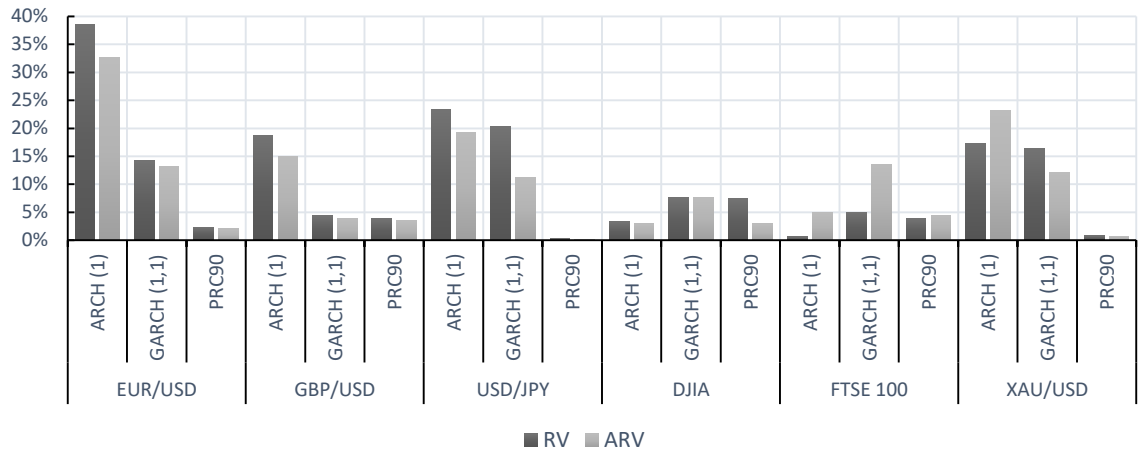
**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five minutes squared returns as in Eq.s (5.1 and 5.2). The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MAE_1$ .

**Figure D.12: Conditional Variance Survival Proportion Dynamics Across the Markets for  $MAE_2$**



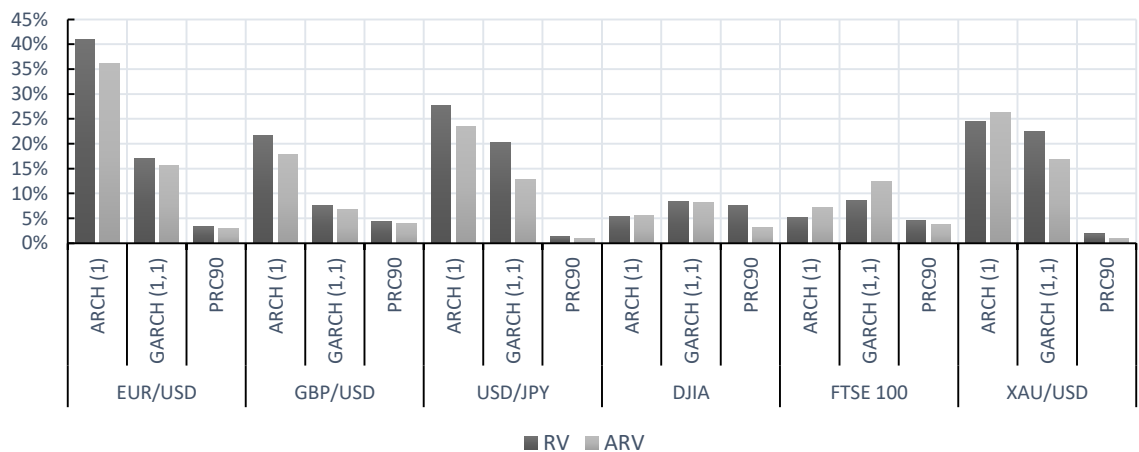
**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five minutes squared returns as in Eq.s (5.1 and 5.2). The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MAE_2$ .

**Figure D.13: Conditional Variance Survival Proportion Dynamics Across the Markets for  $MSE_2$**



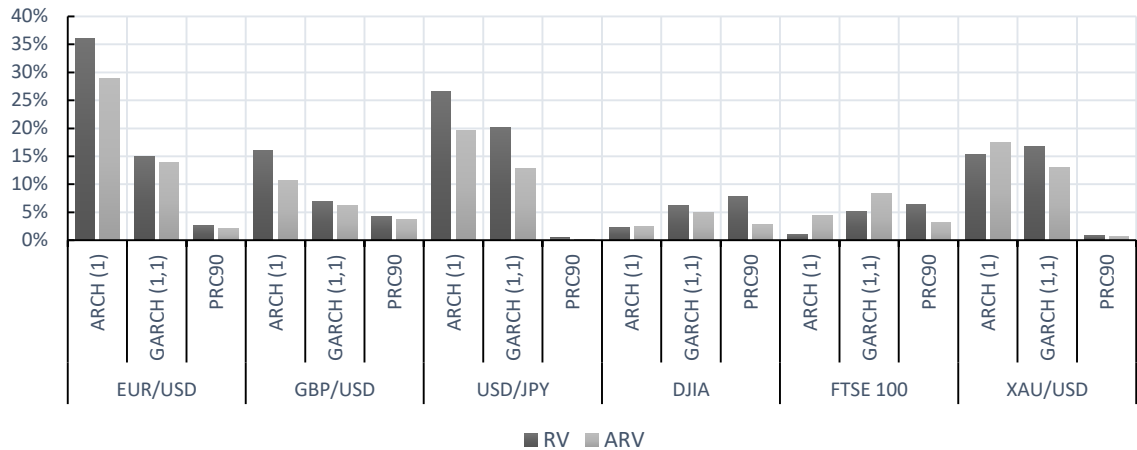
**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five minutes squared returns as in Eq.s (5.1 and 5.2). The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_2$ .

**Figure D.14: Conditional Variance Survival Proportion Dynamics Across the Markets for  $R^2_{LOG}$**



**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five minutes squared returns as in Eq.s (5.1 and 5.2). The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $R^2_{LOG}$ .

**Figure D.15: Conditional Variance Survival Proportion Dynamics Across the Markets for QLIKE**



**Note:** The figure presents the average success rate of alternative variance specifications. The RV and ARV correspond to the realized variance and the adjusted realized variance based on five minutes squared returns as in Eq.s (5.1 and 5.2). The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is QLIKE.

## D.8 Class Analysis for Other Loss Functions

This Appendix investigates the success rate of each of twenty classes of volatility forecasting models. The formulation for all classes of the forecasting model are presented in Eq.s (5.3 to 5.20) of Table 5.1. The Tables D.12 to D.16 present the proportion of the models from each class that survive the tests. The values correspond to averages over the whole study period (2013-2017). The test survivors are based on six loss functions. The results for the  $MSE_1$  are presented in the main text. The other five loss functions provided here are  $MAE_1$ ,  $MAE_2$ ,  $MSE_2$ ,  $R^2LOG$ , and QLIKE. The specifications of the loss functions are provided in Table 5.3.

[Tables D.12 to D.16]

## Bibliography

- Akkoç, S., 2012. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 222(1), pp.168-178.
- Allen, F. and Karjalainen, R. 1999. Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, 51(2), pp. 245-271.
- Alvarez-Diaz, M. and Alvarez, A. 2003. Forecasting exchange rates using genetic algorithms. *Applied Economics Letters*, 10(6), pp.319-322.
- Andersen, T.G. and Bollerslev, T., 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, pp.885-905.
- Andersen, T.G., Bollerslev, T. and Meddahi, N., 2005. Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, 73(1), pp.279-296.
- Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P., 2003. Modeling and forecasting realized volatility. *Econometrica*, 71(2), pp.579-625.
- Andersson, P., Memmert, D. and Popowicz, E. 2009. Forecasting outcomes of the World Cup 2006 in football: Performance and confidence of bettors and laypeople. *Psychology of Sport and Exercise*, 10(1), 116-123.
- Angelini, G. and De Angelis, L., 2017. PARX model for football match predictions. *Journal of Forecasting*, 36(7), pp.795-807.
- Angelov, P.P. and Filev, D.P., 2004. An approach to online identification of Takagi-Sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1), pp.484-498.
- Asai, M. and McAleer, M., 2011. Alternative asymmetric stochastic volatility models. *Econometric Reviews*, 30(5), pp.548-564.

Audas, R., Dobson, S. and Goddard, J., 2002. The impact of managerial change on team performance in professional sports. *Journal of Economics and Business*, 54(6), pp.633-650.

Aye, G., Gupta, R., Hammoudeh, S. and Kim, W.J., 2015. Forecasting the price of gold using dynamic model averaging. *International Review of Financial Analysis*, 41, pp.257-266.

Baboota, R. and Kaur, H. 2018. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, (forthcoming)

Baillie, R.T. and Bollerslev, T., 1989. Common stochastic trends in a system of exchange rates. *The Journal of Finance*, 44(1), pp.167-181.

Baillie, R.T., Bollerslev, T. and Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 74(1), pp.3-30.

Baio, G. and Blangiardo, M., 2010. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2), pp.253-264.

Bajgrowicz, P. and Scaillet, O., 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3), pp.473-491.

Bancroft, T., Du, C. and Nettleton, D., 2013. Estimation of False Discovery Rate Using Sequential Permutation p-Values. *Biometrics*, 69(1), pp.1-7.

Bao, Y., Lee, T.H. and Saltoglu, B., 2006. Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *Journal of forecasting*, 25(2), pp.101-128.

Barras, L., Scaillet, O. and Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance*, 65(1), pp.179-216.

Barras, L., Scaillet, O., & Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance*, 65(1), 179-216.

Bastos, L.S. and da Rosa, J.M.C., 2013. Predicting probabilities for the 2010 FIFA World Cup games using a Poisson-Gamma model. *Journal of Applied Statistics*, 40(7), pp.1533-1544.

Bekiros, S. D., 2010. Heterogeneous trading strategies with adaptive fuzzy actor-critic reinforcement learning: A behavioral approach. *Journal of Economic Dynamics and Control*, 34(6), pp.1153-1170.

Bellman, R.E. and Zadeh, L.A., 1970. Decision-making in a fuzzy environment. *Management Science*, 17(4), pp. B141-B164.

Bena, J., Ferreira, M.A., Matos, P. and Pires, P., 2017. Are foreign investors locusts? The long-term effects of foreign institutional ownership. *Journal of Financial Economics*, 126(1), pp.122-146.

Benavides, G. and Capistrán, C., 2012. Forecasting exchange rate volatility: The superior performance of conditional combinations of time series and option implied forecasts. *Journal of Empirical Finance*, 19(5), pp.627-639.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 72(4), pp.405-416.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pp.289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3), pp.191-203.

BIS, 2013. Triennial Central Bank Survey of foreign exchange and derivatives market activity in 2013, Bank for International Settlement, online report.

Blume, L., Easley, D. and O'hara, M., 1994. Market statistics and technical analysis: The role of volume. *The Journal of Finance*, 49(1), pp.153-181.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), pp.307-327.

Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pp.542-547.

Bollerslev, T., Patton, A.J. and Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), pp.1-18.

Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, pp.3-62.

Boshnakov, G., Kharrat, T. and McHale, I. G. 2017. A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2), 458-466.

Boston Consulting Group, 2015. Where machines could replace humans—and where they can't (yet), Boston Consulting Group, online report

Brock, W., Lakonishok, J. and LeBaron, B., 1992. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of finance*, 47(5), pp.1731-1764.

Brooks, C. and Persaud, G., 2003. Volatility forecasting for risk management. *Journal of forecasting*, 22(1), pp.1-22.

Broto, C. and Ruiz, E., 2004. Estimation methods for stochastic volatility models: a survey. *Journal of Economic Surveys*, 18(5), pp.613-649.



- Burden, F., & Winkler, D., (2009). Bayesian regularization of neural networks. *Artificial Neural Networks: Methods and Applications*, 23-42.
- Byrne, J. P., Korobilis, D. and Ribeiro, P. J., 2016. Exchange rate predictability in a changing world. *Journal of International Money and Finance*, 62, 1-24.
- Cain, M., Law, D. and Peel, D., 2000. The Favourite-Longshot Bias and Market Efficiency in UK Football betting. *Scottish Journal of Political Economy*, 47(1), pp.25-36.
- Candela, J. Q. and Hansen, L. K., 2004. Learning with uncertainty-Gaussian processes and relevance vector machines, Doctoral Dissertation,
- Cesari, R. and Cremonini, D., 2003. Benchmarking, portfolio insurance and technical analysis: a Monte Carlo comparison of dynamic strategies of asset allocation. *Journal of Economic Dynamics and Control*, 27(6), pp.987-1011.
- Chan, J.C. and Grant, A.L., 2016. Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics*, 54, pp.182-189.
- Chang, F.J. and Chang, Y.T., 2006. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Advances in Water Resources*, 29(1), pp.1-10.
- Chang, P.C., Liu, C.H. and Lai, R.K., 2008. A fuzzy case-based reasoning model for sales forecasting in print circuit board industries. *Expert Systems with Applications*, 34(3), pp.2049-2058.
- Cheng, C.B. and Lee, E.S., 2001. Switching regression analysis by fuzzy adaptive network. *European Journal of Operational Research*, 128(3), pp.647-663.
- Chinn, M. D. and Meese, R. A., 1995. Banking on currency forecasts: How predictable is change in money?. *Journal of International Economics*, 38(1), pp.161-178.
- Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. *Journal of Intelligent & fuzzy systems*, 2(3), pp.267-278.

Christoffersen, P.F. and Diebold, F.X., 2000. How relevant is volatility forecasting for financial risk management?. *Review of Economics and Statistics*, 82(1), pp.12-22.

Chui, M., Manyika, J. and Miremadi, M., 2016. Where machines could replace humans—and where they can't (yet), *Mckinsey Quarterly*

Cialenco, I. and Protopapadakis, A., 2011. Do technical trading profits remain in the foreign exchange market? Evidence from 14 currencies. *Journal of International Financial Markets, Institutions and Money*, 21(2), pp.176-206.

Cialenco, I. and Protopapadakis, A., 2011. Do technical trading profits remain in the foreign exchange market? Evidence from 14 currencies. *Journal of International Financial Markets, Institutions and Money*, 21(2), pp.176-206.

Constantinou, A.C., Fenton, N.E. and Neil, M., 2012. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36, pp.322-339.

Corsi, F., Mittnik, S., Pigorsch, C. and Pigorsch, U., 2008. The volatility of realized volatility. *Econometric Reviews*, 27(1-3), pp.46-78.

Crowder, M., Dixon, M., Ledford, A. and Robinson, M., 2002. Dynamic modelling and prediction of English Football League matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(2), pp.157-168.

Demirgüç-Kunt, A. and Levine, R., 1996. Stock market development and financial intermediaries: stylized facts. *The World Bank Economic Review*, 10(2), pp.291-321.

Dickey, D.A. and Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a), pp.427-431.

Dimson, E. and Marsh, P., 1990. Volatility forecasting without data-snooping. *Journal of Banking & Finance*, 14(2-3), pp.399-421.

Dixon, M. and Robinson, M. 1998. A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3), 523-538.

Dixon, M. J. and Pope, P. F. 2004. The value of statistical forecasts in the UK association football betting market. *International journal of forecasting*, 20(4), 697-711.

Dixon, M.J. and Coles, S.G., 1997. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2), pp.265-280.

Dobson, S. and Goddard, J., 2003. Persistence in sequences of football match results: A Monte Carlo analysis. *European Journal of Operational Research*, 148(2), pp.247-256.

Edwards, R. D., Magee, J. and Bassetti, W. C. 2007. Technical analysis of stock trends. CRC Press.

Engle, R.F. and Bollerslev, T., 1986. Modelling the persistence of conditional variances. *Econometric reviews*, 5(1), pp.1-50.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp.987-1007.

Esposito, F.P. and Cummins, M., 2016. Multiple hypothesis testing of market risk forecasting models. *Journal of Forecasting*, 35(5), pp.381-399.

Eurex, 2018. Contract Specifications for Futures Contracts and Options Contracts at Eurex Deutschland and Eurex Zürich. [Online]

Fama, E. F. (1965). The behaviour of stock-market prices. *The Journal of Business*, 38(1), pp. 34-105.

Fama, Eugene F., and Kenneth R. French, 2010, Luck Versus Skill in the Cross-Section of Mutual Fund Returns, *Journal of Finance* 65, 1915-1947.

- Fan, J., & Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 1143-1164.
- Fang, J., Jacobsen, B. and Qin, Y., 2014. Predictability of the simple technical trading rules: An out-of-sample test. *Review of Financial Economics*, 23(1), pp.30-45.
- Feess, E., Müller, H. and Schumacher, C. 2016. Estimating risk preferences of bettors with different bet sizes. *European Journal of Operational Research*, 249(3), 1102-1112.
- Fernández-Rodríguez, F., González-Martel, C and Sosvilla-Rivero, S., S. 2000. On the profitability of technical trading rules based on artificial neural networks:: Evidence from the Madrid stock market. *Economics letters*, 69(1), pp.89-94.
- Figlewski, S., 1997. Forecasting volatility. *Financial markets, institutions & instruments*, 6(1), pp.1-88.
- Fletcher, T., Redpath, F. and D'Alessandro, J., 2009. Machine learning in FX carry basket prediction. In *Proceedings of the World Congress on Engineering* (Vol. 2).
- Forrest, D. and Simmons, R., 2008. Sentiment in the betting market on Spanish football. *Applied Economics*, 40(1), pp.119-126.
- Forrest, D., Goddard, J., and Simmons, R., 2005. Odds-setters as forecasters: The case of English football. *International Journal of forecasting*, 21(3), pp.551-564.
- Gehrig, T. and Menkhoff, L. 2006. Extended evidence on the use of technical analysis in foreign exchange. *International Journal of Finance & Economics*, 11(4), pp. 327-338.
- Gençay, R. 1998. The predictability of security returns with simple technical trading rules. *Journal of Empirical Finance*, 5(4), pp. 347-359.

Gencay, R., Dacorogna, M., Olsen, R. and Pictet, O., 2003. Foreign exchange trading models and market behaviour. *Journal of Economic Dynamics and Control*, 27(6), pp.909-935.

Genovese, C. R., & Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476), 1408-1417.

Ghosh, S. and Mujumdar, P.P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in water resources*, 31(1), pp.132-146.

Goddard, J. and Asimakopoulos, I., 2004. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1), pp.51-66.

Goddard, J., 2005. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, 21(2), pp.331-340.

Goetzmann, W., Ingersoll, J., Spiegel, M. and Welch, I., 2007. Portfolio performance manipulation and manipulation-proof performance measures. *The Review of Financial Studies*, 20(5), pp.1503-1546.

Gomes, J., Portela, F. and Santos, M.F., 2016. Pervasive Decision Support to predict football corners and goals by means of data mining. In *New Advances in Information Systems and Technologies* (pp. 547-556). Springer, Cham.

Gradojevic, N. and Gençay, R., 2013. Fuzzy logic, trading uncertainty and technical trading. *Journal of Banking & Finance*, 37(2), pp.578-586.

Gradojevic, N., 2007. Non-linear, hybrid exchange rate modelling and trading profitability in the foreign exchange market. *Journal of Economic Dynamics and Control*, 31(2), pp.557-574.

Graham, I. and Stott, H., 2008. Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, 40(1), pp.99-109.

Gramacy, R., Malone, S. W. and Horst, E. T. 2014. Exchange rate fundamentals, forecasting, and speculation: Bayesian models in black markets. *Journal of Applied Econometrics*, 29(1), pp.22-41.

Guillaume, S., 2001. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE transactions on fuzzy systems*, 9(3), pp.426-443.

Gustafson, D.E. and Kessel, W.C., 1979. Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes* (pp. 761-766). IEEE.

Hansen, B.E., 1994. Autoregressive conditional density estimation. *International Economic Review*, pp.705-730.

Hansen, P. and Lunde, A., 2011. Forecasting volatility using high frequency data. *The Oxford Handbook of Economic Forecasting*, Oxford: Blackwell, pp.525-556.

Hansen, P. R. 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), pp. 365-380.

Hansen, P. R. and Lunde, A. 2005. A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. *Journal of applied econometrics*, 20(7), pp. 873-889.

Hansen, P.R. and Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. *Journal of applied econometrics*, 20(7), pp.873-889.

Hansen, P.R., 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), pp.365-380.

Harvey, A. and Sucarrat, G., 2014. EGARCH models with fat tails, skewness and leverage. *Computational Statistics & Data Analysis*, 76, pp.320-338.

Harvey, C.R. and Liu, Y., 2015. Backtesting. *The Journal of Portfolio Management*, 42(1), pp.13-28.

Harvey, C.R. and Whaley, R.E., 1992. Market volatility prediction and the efficiency of the S & P 100 index option market. *Journal of Financial Economics*, 31(1), pp.43-73.

Hawking, S., 2014. *Stephen Hawking warns artificial intelligence could end mankind* [Interview] (4 December 2014).

Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B. and Tomasello, M., 2007. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843), pp.1360-1366.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp.65-70.

Hruschka, H., 1988. Use of fuzzy relations in rule-based decision support systems for business planning problems. *European journal of operational research*, 34(3), pp.326-335.

Hsu, P. H. and Kuan, C. M. 2005. Reexamining the profitability of technical analysis with data snooping checks. *Journal of Financial Econometrics*, 3(4), pp. 606-628.

Hsu, P.H., Hsu, Y.C. and Kuan, C.M., 2010. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3), pp.471-484.

Hsu, Y., Kuan, C. and Yen, M., 2014. A generalized stepwise procedure with improved power for multiple inequalities testing. *Journal of Financial Econometrics*, 12(4), pp. 730-755.

Huang, C., Gong, X., Chen, X. and Wen, F., 2013. Measuring and forecasting volatility in Chinese stock market using HAR-CJ-M model. In *Abstract and Applied Analysis* (Vol. 2013). Hindawi.

Huang, S.C., Chuang, P.J., Wu, C.F and Lai, H.J. 2010 Chaos-based support vector regressions for exchange rate forecasting, *Expert Systems with Applications*, 37(12), pp. 8590-8598.

Igiri, C. P. 2015. Support Vector Machine–Based Prediction System for a Football Match Result. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17(3), pp. 21-26.

Investment Technology Group, 2013. Trade Cost and Trade Flow. [Online]

Jang, J.S., 1993. ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3), pp.665-685.

Jasic, T. and Wood, D. 2004. The profitability of daily stock market indices trades based on neural network predictions: Case study for the S&P 500, the DAX, the TOPIX and the FTSE in the period 1965-1999. *Applied Financial Economics*, 14(4), pp.285-297.

Jorion, P., 1996. Risk and turnover in the foreign exchange market. In Frankel, J., Galli, G. and Giovannini, A. (eds) *The microstructure of foreign exchange markets*, Chicago, Ill.: University of Chicago Press, pp.19-40.

Joseph, A., Fenton, N.E. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7), pp.544-553.

Junior, M.V.W. and Pereira, P.L.V., 2011. Modeling and Forecasting of Realized Volatility: Evidence from Brazil. *Brazilian Review of Econometrics*, 31(2), pp.315-337.

Kambouroudis, D.S., McMillan, D.G. and Tsakou, K., 2016. Forecasting stock return volatility: A comparison of garch, implied volatility, and realized volatility models. *Journal of Futures Markets*, 36(12), pp.1127-1163.

Kang, S.H., Kang, S.M. and Yoon, S.M., 2009. Forecasting volatility of crude oil markets. *Energy Economics*, 31(1), pp.119-125.



- Karlis, D. and Ntzoufras, I., 2008. Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2), pp.133-145.
- Kelly, J.L., 1956. A new interpretation of information rate. *Bell Labs Technical Journal*, 35(4), pp.917-926.
- Kilian, L. and Taylor, M.P., 2003. Why is it so difficult to beat the random walk forecast of exchange rates?. *Journal of International Economics*, 60(1), pp.85-107.
- Koning, R.H., 2000. Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3), pp.419-431.
- Kontonikas, A., MacDonald, R. and Saggu, A., 2013. Stock market reaction to fed funds rate surprises: State dependence and the financial crisis. *Journal of Banking & Finance*, 37(11), pp.4025-4037.
- Koop, G. and Korobilis, D., 2012. Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3), pp.867-886.
- Koopman, S.J. and Lit, R., 2015. A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), pp.167-186.
- Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis, *Journal of Finance* 61, 2551-2595.
- Kulinskaya, E. and Lewin, A., 2009. On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96(1), pp.201-211.
- Kuncheva, L.I., 2000. How good are fuzzy if-then classifiers?. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 30(4), pp.501-509.

- Kuo, R.J., 2001. A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research*, 129(3), pp.496-517.
- Kuypers, T., 2000. Information and efficiency: an empirical study of a fixed odds betting market. *Applied Economics*, 32(11), pp.1353-1363.
- Li, J. and Xiu, D., 2016. Generalized Method of Integrated Moments for High-Frequency Data. *Econometrica*, 84(4), pp.1613-1633.
- Liang, K. (2016). False discovery rate estimation for large-scale homogeneous discrete p-values. *Biometrics*, 72(2), 639-648.
- Liang, K. and Nettleton, D., 2012. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1), pp.163-182.
- Liang, K., 2016. False discovery rate estimation for large-scale homogeneous discrete p-values. *Biometrics*, 72(2), pp.639-648.
- Lin, K.P. and Pai, P.F. 2010. A fuzzy support vector regression model for business cycle predictions, *Expert Systems with Applications*, 37 (7), pp. 5430-5435.
- Lo, A.W., 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, 30, pp. 15-29.
- Lopez, J.A., 2001. Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, 20(2), pp.87-109.
- MacDonald, R. and Taylor, M.P., 1994. The monetary model of the exchange rate: long-run relationships, short-run dynamics and how to beat a random walk. *Journal of International Money and finance*, 13(3), pp.276-290.
- MacKay, D.J., 1992. Bayesian interpolation. *Neural Computation*, 4(3), pp.415-447.

Maclean, L.C., Thorp, E.O. and Ziemba, W.T., 2010. Long-term capital growth: the good and bad properties of the Kelly and fractional Kelly capital growth criteria. *Quantitative Finance*, 10(7), pp.681-687.

Markram, H., 2012. The human brain project. *Scientific American*, 306(6), pp.50-55.

Martens, D., Baesens, B., Van Gestel, T. and Vanthienen, J., 2007. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3), pp.1466-1476.

Martins, R. G., Martins, A. S., Neves, L. A., Lima, L. V., Flores, E. L., and do Nascimento, M. Z. 2017. Exploring polynomial classifier to predict match results in football championships. *Expert Systems with Applications*, 83, 79-93.

Mathur, N., Glesk, I. and Buis, A., 2016. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian processes for machine learning (GPML) algorithms for the prediction of skin temperature in lower limb prostheses. *Medical Engineering & Physics*, 38(10), pp.1083-1089.

Mazzocco, M. and Saini, S., 2012. Testing efficient risk sharing with heterogeneous risk preferences. *The American Economic Review*, 102(1), pp.428-468.

Meeden, G., 1981. Betting Against a Bayesian Bookie. *Journal of the American Statistical Association*, 76(373), pp. 202-204.

Meese, R.A. and Rogoff, K., 1983. Empirical exchange rate models of the seventies: Do they fit out of sample?. *Journal of international economics*, 14(1), pp.3-24.

Min, B., Kim, J., Choe, C., Eom, H. and McKay, R.B., 2008. A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, 21(7), pp.551-562.

MSCI, 2013. Deploying Multi-Factor Index Allocations in Institutional Portfolios. [Online]

Neely, C., Weller, P. and Dittmar, R. 1997. Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of Financial and Quantitative Analysis*, 32(4), pp.405-426.

Neely, C.J. and Weller, P.A., 2013. Lessons from the evolution of foreign exchange trading strategies. *Journal of Banking & Finance*, 37(10), pp.3783-3798.

Neely, C.J., Weller, P.A. and Ulrich, J.M., 2009. The adaptive markets hypothesis: evidence from the foreign exchange market. *Journal of Financial and Quantitative Analysis*, 44(2), pp.467-488.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pp.347-370.

Newall, P. (2013). Further limit hold 'em: Exploring the model poker game. Las Vegas, Nevada: Two Plus Two Publishing.

Newall, P.W., 2017. Behavioral complexity of British gambling advertising. *Addiction Research & Theory*, 25(6), pp.505-511.

Newall, P.W., 2018. Commentary: Heads-up limit hold 'em poker is solved. *Frontiers in Psychology*, 9, p.210.

Oberstone, J. 2009. Differentiating the top English premier league football clubs from the rest of the pack: Identifying the keys to success. *Journal of Quantitative Analysis in Sports*, 5(3).

Oberstone, J. 2011. Comparing team performance of the English premier league, Serie A, and La Liga for the 2008-2009 season. *Journal of Quantitative Analysis in Sports*, 7(1).

Owramipur, F., Eskandarian, P. and Mozneb, F.S., 2013. Football result prediction with Bayesian network in Spanish League-Barcelona team. *International Journal of Computer Theory and Engineering*, 5(5), p.812.

Pai, P.F., Lin, C.S., Hong, W.C. and Chen, C.T. 2006 A hybrid support vector machine regression for exchange rate prediction, *International Journal of Information and Management Sciences*, 17 (2), pp. 19-32.

Park, J. and Sandberg, I.W., 1991. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2), pp.246-257

Pesaran, M.H. and Timmermann, A., 1992. A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4), pp.461-465.

Phillips, P.C. and Perron, P., 1988. Testing for a unit root in time series regression. *Biometrika*, 75(2), pp.335-346.

Piramuthu, S., 1999. Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, 112(2), pp.310-321.

Polat, K. and Güneş, S., 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), pp.702-710.

Politis, D.N. and Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical association*, 89(428), pp.1303-1313. Poon, S.H. and Granger, C.W., 2003. Forecasting volatility in financial markets: A review. *Journal of economic literature*, 41(2), pp.478-539.

Pounds, S. and Cheng, C., 2006. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16), pp.1979-1987.

Psaradellis, I., Laws, J., Pantelous, A., Sermpinis, S., (2017). Technical Trading, False Discoveries & Familywise Errors : The Case of Crude Oil. Working Paper.

PwC. 2017. UK Economic Outlook March 2017. [ONLINE] Available at: <https://www.pwc.co.uk/economic-services/ukeyo/pwcukkeyo-section-4-automation-march-2017-v2.pdf>. [Accessed 17 April 2018].

Qi, M. and Wu, Y., 2006. Technical trading-rule profitability, data snooping, and reality check: evidence from the foreign exchange market. *Journal of Money, Credit, and Banking*, 38(8), pp.2135-2158.

Raftery, A.E., Kárný, M. and Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1), pp.52-66.

Rime, D., Sarno, L. and Sojli, E., 2010. Exchange rate forecasting, order flow and macroeconomic information. *Journal of International Economics*, 80(1), pp.72-88.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237-1282.

Romano, J. P., Shaikh, A. M., & Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2), 404-447.

Romano, J.P. and Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), pp.1237-1282.

Romano, J.P. and Wolf, M., 2007. Control of generalized error rates in multiple testing. *The Annals of Statistics*, pp.1378-1408.

Romano, J.P. and Wolf, M., 2010. Balanced control of generalized error rates. *The Annals of Statistics*, pp.598-633.

Romano, J.P., Shaikh, A.M. and Wolf, M., 2008. Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2), pp.404-447.

Rotshtein, A. P., Posner, M., and Rakityanskaya, A. B. 2005. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4), 619-630.

Rue, H. and Salvesen, O., 2000. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3), pp.399-418.

- Sadorsky, P., 2005. Stochastic volatility forecasting and risk management. *Applied Financial Economics*, 15(2), pp.121-135.
- Schumaker, R. P., Jarmoszko, A. T. and Labeledz Jr, C. S. 2016. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, 88, 76-84.
- Sermpinis, G., Laws, J. and Dunis, C.L., 2015. Modelling commodity value at risk with Psi Sigma neural networks using open-high-low-close data. *The European Journal of Finance*, 21(4), pp.316-336.
- Sermpinis, G., Theofilatos, K.A, Karathanasopoulos, A.S., Georgopoulos, E.F. and Dunis, C.L. 2013. Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and Particle Swarm Optimization, *European Journal of Operational Research*, 225 (3), pp. 528-540.
- Shapiro, A. F. (2000), A Hitchhiker's Guide to the Techniques of Adaptive Nonlinear Models, *Insurance, Mathematics and Economics*, 26(2), pp. 119-132.
- Shapiro, A.F., 2002. The merging of neural networks, fuzzy logic, and genetic algorithms. *Insurance: Mathematics and Economics*, 31(1), pp.115-131.
- Singpurwalla, N.D. and Booker, J.M., 2004. Membership functions and probability measures of fuzzy sets. *Journal of the American Statistical Association*, 99(467), pp.867-877.
- Sortino, F.A. and Price, L.N., 1994. Performance measurement in a downside risk framework. *Journal of Investing*, 3(3), pp.59-64.
- Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6), 2013-2035.
- Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), 187-205.

Storey, J. D., Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100, 9440-9445.

Storey, J.D., 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), pp.479-498.

Storey, J.D., Taylor, J.E. and Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), pp.187-205.

Štrumbelj, E. 2014. On determining probability forecasts from betting odds. *International journal of forecasting*, 30(4), 934-943.

Štrumbelj, E. and Šikonja, M. R. 2010. Online bookmakers' odds as forecasts: The case of European soccer leagues. *International Journal of Forecasting*, 26(3), 482-488.

Sugeno, M., 1985. Industrial applications of fuzzy control. *Elsevier Science Inc., New York, NY, USA*.

Sullivan, R., Timmermann, A. and White, H., 1999. Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, 54(5), pp.1647-1691.

Sun, W., Reich, B. J., Tony Cai, T., Guindani, M., & Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1), 59-83.

Sweeney, R. J. 1988. Some new filter rule tests: Methods and results. *Journal of Financial and Quantitative Analysis*, 23(03), pp. 285-300.

Taylor, M.P. 1992. The use of technical analysis in the foreign exchange market. *Journal of International Money and Finance*, 11(3), pp.304-314.

Taylor, N. (2014). The rise and fall of technical trading rule success. *Journal of Banking & Finance*, 40, 286-302.



- Taylor, S.J., 1982. Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, 1, pp.203-226.
- Teodorović, D., 1994. Fuzzy sets theory applications in traffic and transportation. *European Journal of Operational Research*, 74(3), pp.379-390.
- Thorp, E. O. (2008). The Kelly criterion in blackjack sports betting, and the stock market. In Zenios, S.A and Ziemba, W.T. (Eds) *Handbook of asset and liability management*, 385-428.
- Ticknor, J. L. 2013. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, 40(14), pp. 5501-5506.
- Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun), pp.211-244.
- Trawinski, K., 2010, A fuzzy classification system for prediction of the results of the basketball games. In *IEEE International Conference on Fuzzy Systems (FUZZ)*, 2010 (pp. 1-7). IEEE.
- Vapnik, V.N., 1998. *Statistical learning theory*. New York: Wiley.
- Vlastakis, N., Dotsis, G. and Markellos, R.N., 2008. Nonlinear modelling of European football scores using support vector machines. *Applied Economics*, 40(1), pp.111-118.
- Vlastakis, N., Dotsis, G. and Markellos, R.N., 2009. How efficient is the European football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28(5), pp.426-444.
- Wang, L.X. and Mendel, J.M., 1992. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE Transactions on Neural Networks*, 3(5), pp.807-814.

- Wang, Y., Pan, Z. and Wu, C., 2018. Volatility spillover from the US to international stock markets: A heterogeneous volatility spillover GARCH model. *Journal of Forecasting*, 37(3), pp.385-400.
- Wei, Y., Wang, Y. and Huang, D., 2010. Forecasting crude oil market volatility: Further evidence using GARCH-class models. *Energy Economics*, 32(6), pp.1477-1484.
- White, H., 2000. A reality check for data snooping. *Econometrica*, 68(5), pp.1097-1126.
- Yan, Xuemin (Sterling), and Lingling Zheng, 2016, Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach, forthcoming Review of Financial Studies.
- Yang, J., Su, X. and Kolari, J.W., 2008. Do Euro exchange rates follow a martingale? Some out-of-sample evidence. *Journal of Banking & Finance*, 32(5), pp.729-740.
- Yegnanarayana, B. 2009. *Artificial neural networks*. PHI Learning Pvt. Ltd.
- Yilmaz, K., 2003. Martingale Property of Exchange Rates and Central Bank Interventions. *Journal of Business & Economic Statistics*, pp.383-395.
- Yu, J., 2002. Forecasting volatility in the New Zealand stock market. *Applied Financial Economics*, 12(3), pp.193-202.
- Zadeh, L.A., 1965. Fuzzy sets. *Information and control*, 8(3), pp.338-353.
- Zadeh, L.A., 1983. The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy sets and systems*, 11(1), pp.199-227.
- Zakoian, J.M., 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, 18(5), pp.931-955.
- Zhang, G., Patuwo, B. E. and Hu, M. Y. 1998. Forecasting with artificial neural networks: The state of the art. *International journal of forecasting*, 14(1), 35-62.

## Tables

**Table 2.1: Descriptive Statistics**

Period	Statistic	EUR/USD	GBP/USD	USD/JPY
(1) 2010.01.04 - 2013.12.31	Mean (bp)	-4.59	-0.28	1.25
	Standard Deviation (bp)	62.2	50.9	63.1
	Kurtosis	3.78	3.25	8.91
	Skewness	-0.12	-0.07	-0.55
	JB p-value	0.00	0.00	0.00
	ADF p-value	0.00	0.00	0.00
(2) 2011.01.03 - 2014.12.31	Mean (bp)	-0.97	-0.02	0.38
	Standard Deviation (bp)	53.8	43.9	59.3
	Kurtosis	4.28	3.51	9.27
	Skewness	-0.14	-0.11	0.06
	JB p-value	0.00	0.00	0.00
	ADF p-value	0.00	0.00	0.00
(3) 2012.01.02 - 2015.12.31	Mean (bp)	-1.68	-0.13	4.30
	Standard Deviation (bp)	54.5	45.5	56.7
	Kurtosis	4.95	3.96	6.51
	Skewness	0.17	0.17	0.05
	JB p-value	0.00	0.00	0.00
	ADF p-value	0.00	0.00	0.00
(4) 2013.01.02 - 2016.12.30	Mean (bp)	-2.19	-0.04	0.29
	Standard Deviation (bp)	54.7	62.1	64.8
	Kurtosis	5.48	4.47	6.81
	Skewness	0.11	-3.42	-0.27
	JB p-value	0.00	0.00	0.00
	ADF p-value	0.00	0.00	0.00

**Note:** The mean and standard deviations are reported in basis points. Reported values of zero for the JB and ADF (without any lagged difference) tests correspond to p-values less than 1 over 100 ( $p < 0.01$ )

Table 2.2: Trading Rules Excess Annualized Return and Sharpe Ratio

Exercise	Asset	IS Performance	OOS Performance	Surviving Rules Count
2013	EUR/USD	3.07% (0.29)	-1.23% (-0.72)	839
	GBP/USD	1.1% (0.29)	0.21% (0.05)	788
	USD/JPY	3.24% (0.07)	-3.37% (-0.65)	870
2014	EUR/USD	5.71% (1.01)	3.69% (0.96)	819
	GBP/USD	5.67% (2.62)	4.41% (1.56)	368
	USD/JPY	6.7% (0.58)	6.19% (1.07)	439
2015	EUR/USD	3.18% (1.45)	2.21% (0.41)	1144
	GBP/USD	6.42% (1.85)	1.47% (0.27)	1047
	USD/JPY	12.5% (2.13)	-6.58% (-0.94)	523
2016	EUR/USD	10.31% (0.25)	2.04% (0.41)	1147
	GBP/USD	4.85% (0.14)	1.91% (0.22)	981
	USD/JPY	8.77% (1.43)	-1.31% (-0.33)	337
<b>Total</b>	Average	5.96% (1.01)	0.8% (0.19)	775.17

Note: The table presents the excess annualized returns (above the risk-free rate) of the technical rules after transaction costs. The values in parentheses correspond to the Sharpe ratios. Trading rules correspond to the number of genuine trading rules identified in the IS by the Romano *et al.* (2008) test combined with the balancing procedure of Romano and Wolf (2010).

Table 2.3: EUR/USD Trading Performance – Annualized Return

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	-2.88%	-0.60%	-0.09%	3	SA (5)	0.26%	1.24%	-2.37%
	SA (10)	1.17%	-2.11%	-0.19%		SA (10)	-2.18%	-1.46%	-3.12%
	SA (15)	-1.44%	-0.68%	-2.22%		SA (15)	-0.99%	-1.08%	-0.64%
	NB (5)	1.06%	0.63%	1.10%		NB (5)	4.23%	2.14%	2.27%
	NB (10)	2.09%	1.58%	2.18%		NB (10)	0.86%	1.52%	2.72%
	NB (15)	1.80%	1.31%	1.82%		NB (15)	1.18%	1.74%	3.07%
	DMA (5)	4.34%	4.55%	4.99%		DMA (5)	<b>7.79%</b>	4.95%	4.59%
	DMA (10)	<b>6.89%</b>	4.26%	<b>6.02%</b>		DMA (10)	6.85%	4.54%	6.98%
	DMA (15)	5.30%	5.49%	5.40%		DMA (15)	6.01%	<b>6.49%</b>	5.23%
	DMS (5)	3.74%	<b>6.86%</b>	4.09%		DMS (5)	6.23%	3.92%	4.30%
	DMS (10)	5.02%	5.06%	5.43%		DMS (10)	6.97%	4.22%	4.44%
	DMS (15)	4.80%	4.99%	2.63%		DMS (15)	6.12%	4.36%	5.86%
	BNN (5)	6.01%	5.91%	4.82%		BNN (5)	5.08%	4.77%	4.82%
	BNN (10)	6.48%	6.51%	5.01%		BNN (10)	6.00%	5.29%	<b>7.27%</b>
	BNN (15)	6.12%	6.00%	5.96%		BNN (15)	5.37%	4.90%	6.08%
2	RVM	4.35%	4.35%	4.35%	4	RVM	4.41%	4.41%	4.41%
	SA (5)	0.27%	-0.70%	1.95%		SA (5)	0.21%	-0.32%	-2.69%
	SA (10)	1.89%	-0.82%	1.60%		SA (10)	-2.46%	1.24%	-3.01%
	SA (15)	0.16%	1.78%	2.26%		SA (15)	-1.32%	1.08%	-1.44%
	NB (5)	1.82%	2.04%	1.11%		NB (5)	2.28%	-2.66%	1.82%
	NB (10)	2.74%	1.79%	3.04%		NB (10)	3.18%	1.80%	1.49%
	NB (15)	1.72%	1.96%	2.00%		NB (15)	3.33%	3.02%	2.77%
	DMA (5)	3.86%	<b>6.97%</b>	7.09%		DMA (5)	6.36%	5.46%	5.14%
	DMA (10)	6.08%	3.82%	6.18%		DMA (10)	5.07%	<b>6.28%</b>	5.80%
	DMA (15)	6.15%	3.43%	5.72%		DMA (15)	6.74%	5.28%	4.14%
	DMS (5)	3.99%	3.26%	5.17%		DMS (5)	4.50%	3.35%	5.02%
	DMS (10)	4.24%	3.14%	6.03%		DMS (10)	5.29%	5.80%	5.93%
	DMS (15)	5.40%	3.25%	6.12%		DMS (15)	6.80%	5.94%	4.01%
	BNN (5)	3.77%	3.01%	6.32%		BNN (5)	6.21%	5.12%	5.57%
	BNN (10)	<b>6.42%</b>	5.38%	5.38%		BNN (10)	6.01%	6.01%	<b>7.44%</b>
	BNN (15)	5.37%	5.04%	<b>7.49%</b>		BNN (15)	<b>7.28%</b>	6.05%	6.19%
	RVM	3.28%	3.28%	3.28%		RVM	4.24%	4.24%	4.24%

**Note:** The table presents the annualized excess return (above the risk-free rate) for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.

Table 2.4: GBP/USD Trading Performance – Annualized Return

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	1.92%	-3.10%	0.08%	3	SA (5)	-0.48%	1.25%	-1.32%
	SA (10)	0.80%	-0.77%	-3.32%		SA (10)	1.66%	-1.92%	2.61%
	SA (15)	-0.53%	-0.05%	-2.48%		SA (15)	1.58%	-0.99%	-2.58%
	NB (5)	2.70%	1.56%	1.10%		NB (5)	1.85%	2.00%	2.52%
	NB (10)	2.43%	2.70%	3.09%		NB (10)	2.55%	3.80%	3.39%
	NB (15)	3.82%	3.78%	1.67%		NB (15)	3.44%	4.09%	1.22%
	DMA (5)	4.57%	4.87%	3.98%		DMA (5)	5.96%	4.17%	4.32%
	DMA (10)	3.89%	5.63%	<b>6.22%</b>		DMA (10)	5.68%	4.79%	5.04%
	DMA (15)	6.66%	5.95%	3.76%		DMA (15)	<b>7.19%</b>	<b>7.24%</b>	4.44%
	DMS (5)	4.60%	4.18%	4.16%		DMS (5)	6.07%	3.05%	3.41%
	DMS (10)	3.92%	5.67%	3.69%		DMS (10)	6.01%	4.42%	5.30%
	DMS (15)	5.20%	5.04%	3.02%		DMS (15)	6.22%	6.92%	5.37%
	BNN (5)	4.89%	5.28%	6.12%		BNN (5)	5.14%	5.35%	5.02%
	BNN (10)	6.38%	5.00%	4.74%		BNN (10)	4.55%	5.84%	3.11%
	BNN (15)	<b>7.29%</b>	<b>6.48%</b>	3.70%		BNN (15)	6.17%	6.90%	<b>6.29%</b>
2	RVM	4.61%	4.61%	4.61%	4	RVM	3.72%	3.72%	3.72%
	SA (5)	0.71%	0.71%	-2.49%		SA (5)	-0.13%	1.99%	0.36%
	SA (10)	2.22%	0.90%	1.00%		SA (10)	1.98%	1.25%	2.58%
	SA (15)	2.40%	0.06%	-0.86%		SA (15)	1.02%	0.34%	-.17%
	NB (5)	2.42%	2.42%	1.94%		NB (5)	2.78%	2.30%	2.79%
	NB (10)	2.62%	3.00%	1.15%		NB (10)	3.05%	3.43%	3.18%
	NB (15)	1.88%	4.04%	0.77%		NB (15)	2.85%	4.02%	3.60%
	DMA (5)	3.64%	3.64%	5.57%		DMA (5)	4.70%	4.87%	5.34%
	DMA (10)	4.03%	3.85%	<b>5.67%</b>		DMA (10)	5.36%	5.61%	<b>5.98%</b>
	DMA (15)	<b>6.53%</b>	5.29%	4.09%		DMA (15)	<b>7.78%</b>	<b>7.40%</b>	5.09%
	DMS (5)	3.00%	3.00%	5.42%		DMS (5)	3.21%	3.91%	4.09%
	DMS (10)	4.48%	3.04%	4.50%		DMS (10)	4.39%	5.20%	4.44%
	DMS (15)	5.57%	5.05%	3.59%		DMS (15)	6.40%	6.17%	4.82%
	BNN (5)	5.16%	4.16%	4.94%		BNN (5)	5.03%	6.02%	5.28%
	BNN (10)	5.21%	<b>5.87%</b>	5.14%		BNN (10)	5.83%	5.95%	5.31%
	BNN (15)	6.29%	5.14%	4.21%		BNN (15)	6.01%	7.54%	3.15%
	RVM	4.19%	4.19%	4.19%		RVM	4.27%	4.27%	4.27%

**Note:** The table presents the annualized excess return (above the risk-free rate) for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.

Table 2.5: USD/JPY Trading Performance – Annualized Return

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	-2.24%	2.72%	3.09%	3	SA (5)	-2.75%	-1.77%	-2.96%
	SA (10)	-0.20%	3.54%	4.07%		SA (10)	-1.72%	-1.45%	1.04%
	SA (15)	-1.16%	0.98%	3.74%		SA (15)	0.21%	-0.93%	-1.52%
	NB (5)	2.44%	3.85%	3.45%		NB (5)	2.09%	2.26%	-1.59%
	NB (10)	2.18%	2.86%	4.24%		NB (10)	3.58%	0.32%	0.30%
	NB (15)	3.07%	2.04%	3.62%		NB (15)	2.77%	2.70%	0.45%
	DMA (5)	3.84%	5.88%	<b>7.22%</b>		DMA (5)	4.42%	4.50%	<b>6.19%</b>
	DMA (10)	4.43%	4.45%	5.51%		DMA (10)	4.67%	4.21%	5.70%
	DMA (15)	5.18%	<b>6.96%</b>	6.29%		DMA (15)	5.40%	6.36%	4.02%
	DMS (5)	5.62%	4.74%	6.65%		DMS (5)	4.83%	3.91%	2.72%
	DMS (10)	5.09%	3.86%	4.93%		DMS (10)	4.79%	3.09%	4.85%
	DMS (15)	4.14%	6.19%	4.30%		DMS (15)	4.91%	6.44%	3.26%
	BNN (5)	4.10%	5.09%	4.84%		BNN (5)	<b>5.99%</b>	5.43%	4.36%
	BNN (10)	5.03%	4.88%	5.47%		BNN (10)	4.85%	4.90%	5.70%
	BNN (15)	<b>6.15%</b>	6.01%	5.70%		BNN (15)	4.23%	<b>6.58%</b>	5.51%
2	RVM	5.22%	5.22%	5.22%	4	RVM	4.01%	4.01%	4.01%
	SA (5)	-2.28%	1.88%	2.31%		SA (5)	-3.33%	-2.63%	-2.79%
	SA (10)	-3.63%	2.46%	1.65%		SA (10)	-2.09%	-2.96%	-1.82%
	SA (15)	0.10%	0.44%	0.80%		SA (15)	0.86%	-0.62%	-3.10%
	NB (5)	2.12%	2.48%	1.15%		NB (5)	1.61%	2.60%	-1.37%
	NB (10)	2.92%	1.37%	2.18%		NB (10)	1.87%	2.15%	2.44%
	NB (15)	3.06%	2.24%	3.24%		NB (15)	2.12%	2.04%	1.28%
	DMA (5)	5.00%	5.40%	<b>6.03%</b>		DMA (5)	<b>7.33%</b>	7.07%	4.69%
	DMA (10)	5.97%	6.76%	5.88%		DMA (10)	5.36%	<b>8.30%</b>	5.91%
	DMA (15)	<b>6.03%</b>	5.54%	3.22%		DMA (15)	6.08%	7.32%	6.54%
	DMS (5)	3.04%	4.45%	2.11%		DMS (5)	5.70%	6.07%	4.63%
	DMS (10)	4.06%	2.75%	4.31%		DMS (10)	4.92%	7.11%	4.82%
	DMS (15)	5.42%	5.91%	5.06%		DMS (15)	5.67%	6.04%	5.28%
	BNN (5)	4.91%	<b>6.36%</b>	3.42%		BNN (5)	6.78%	5.23%	4.55%
	BNN (10)	5.74%	5.06%	5.35%		BNN (10)	6.00%	5.82%	<b>6.71%</b>
	BNN (15)	6.00%	6.12%	5.88%		BNN (15)	6.84%	5.09%	4.90%
	RVM	4.75%	4.75%	4.75%		RVM	5.47%	5.47%	5.47%

**Note:** The table presents the annualized excess return (above the risk-free rate) for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.

Table 3.1: Football Forecasting Literature Comparison

Relevant Studies	Main model	Betting Application	Kelly Criterion	Profitability
<b>This study</b>	Conditional Fuzzy Logic	Yes	Yes	Yes
<b>Audas et al. (2002), Dobson and Goddard (2003), Goddard (2005)</b>	Ordered Probit	N/A	N/A	N/A
<b>Forrest et al. (2005), Graham and Scott (2008)</b>	Ordered Probit	Yes	N/A	N/A
<b>Kuypers (2000), Goddard and Asimakopoulos (2004), Forrest and Simmons (2008)</b>	Ordered Probit	Yes	N/A	Yes
<b>Rotshtein et al. (2005), Trawinski (2010), Bastos and Rosa (2013)</b>	Fuzzy Logic	N/A	N/A	N/A
<b>Meeden (1981), Dixon and Coles (1997), Rue and Salvesen (2000), Vlastakis et al. (2009), Constantinou et al. (2012), Koopman and Lit (2015), Angelini and De Angelis (2017)</b>	Probabilistic Approach	Yes	N/A	Yes
<b>Joseph et al. (2006), Karlis and Ntzoufras (2008), Min et al. (2008), Crowder et al. (2002), Baio and Blangiardo (2010), Owrapipur et al. (2013)</b>	Probabilistic Approach	N/A	N/A	N/A
<b>Vlastakis et al. (2008)</b>	Support Vector Machine	Yes	N/A	Yes
<b>Gomes et al. (2016), Martins et al. (2017), Baboota and Kaur (2018)</b>	Support Vector Machine	N/A	N/A	N/A



Table 3.2: Inputs Series

Points of H team –Points of A team before the start of the game	Points of H in the last 1,2,3 games and from the start of the season	Points of A in the last 1,2,3 games and from the start of the season	Points of H in the last 1,2,3 games and from the start of the season when H plays at home	Points of A in the last 1,2,3 games and from the start of the season when A plays at away
Number of goals of H team in the last 1,2,3 games and from the start of the season	Number of goals of A team in the last 1,2,3 games and from the start of the season	Number of goals of H team in the last 1,2,3 games and from the start of the season when H team plays at home	Number of goals of A team in the last 1,2,3 games and from the start of the season when A team plays away	Number of shots on target of H team in the last 1,2,3 games and from the start of the season
Number of shots on target of H team in the last 1,2,3 games and from the start of the season	Number of shots on target of A team in the last 1,2,3 games and from the start of the season	Number of shots on target of H team in the last 1,2,3 games and from the start of the season when H team plays at home	Number of shots on target of H team in the last 1,2,3 and from the start of the season games when A team plays at away	Number of corner kicks of H team in the last 1,2,3 games and from the start of the season
Number of corner kicks of A team in the last 1,2,3 and from the start of the season games	Number of corner kicks of H team in the last 1,2,3 games and from the start of the season when H team plays at home	Number of corner kicks of A team in the last 1,2,3 games and from the start of the season when A team plays at away	H team booking points in the last 1,2,3 games and from the start of the season	A team booking points in the last 1,2,3 games and from the start of the season
Betbrain average home win odds	Betbrain average draw odds	Betbrain average away win odds	Betbrain average over 2.5 goals odds	Betbrain average under 2.5 goals odds
Betbrain size of handicap (home team)	Betbrain average Asian handicap home team odds	Betbrain average Asian handicap away team odds		

**Notes:** There are six main categories of predictors. The first one is based on the odds offered by the bookies and the other five originate from the performance of the teams over the past games. The inputs categories are: (i) odds for match outcome, number of goals and Asian handicap size bets (home team winning, draw, away team winning, over 2.5 goals, under 2.5 goals, home team win, and away team win Asian handicaps – 8 total), (ii) Points achieved, (iii) Goals scored (iv) Corner kicks, (v) Shots on target, (vi) Booking points. For the last five categories, 15 different types of measures are introduced to make sure all types of developments are considered. Team H is the home team and team A is the away team. Booking points are 25 points per red card and 10 per yellow card in a game. The total number of inputs is 83.

Table 3.3: Accuracy Ratios (Game Result)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	68.75%	75.00%	66.67%	61.54%	72.73%	66.65%	68.56%**
		2	63.60%	50.03%	62.50%	57.14%	67.41%	50.00%	58.45%**
	La -Liga	1	60.00%	48.33%	57.69%	50.04%	55.81%	46.55%	53.07%**
		2	63.60%	43.50%	56.68%	43.47%	44.68%	41.60%	48.92%**
	Seria A	1	53.33%	54.50%	61.53%	60.00%	57.45%	61.50%	58.05%**
		2	53.06%	52.63%	54.56%	66.10%	66.70%	53.57%	57.77%**
RVM	Premiership	1	46.05%	37.84%	38.25%	44.33%	35.02%	52.02%	42.25%**
		2	41.22%	37.58%	42.61%	40.23%	34.66%	46.32%	41.59%**
	La -Liga	1	40.33%	36.00%	39.38%	39.33%	35.67%	41.33%	38.68%**
		2	33.67%	36.99%	43.33%	35.33%	41.00%	37.00%	37.89%*
	Seria A	1	51.35%	50.34%	45.92%	44.96%	40.60%	39.08%	45.38%**
		2	47.97%	44.22%	46.22%	34.90%	39.44%	40.27%	42.17%**
ANFIS	Premiership	1	45.70%	40.20%	35.91%	39.18%	40.67%	39.33%	40.17%**
		2	35.81%	39.93%	37.46%	37.00%	36.67%	41.33%	38.03%**
	La -Liga	1	40.00%	36.67%	41.10%	35.33%	34.33%	40.33%	37.96%*
		2	38.33%	34.59%	41.00%	36.67%	40.00%	36.03%	37.77%*
	Seria A	1	42.91%	35.47%	42.18%	42.44%	40.27%	40.14%	40.57%**
		2	34.80%	41.16%	43.70%	36.24%	38.38%	41.61%	39.31%**
OP	Premiership	1	46.70%	44.93%	43.60%	38.49%	46.67%	47.00%	46.69%**
		2	45.61%	47.65%	41.90%	43.33%	39.66%	39.32%	41.53%**
	La -Liga	1	47.67%	52.67%	52.40%	48.67%	42.67%	51.67%	49.29%**
		2	42.60%	49.66%	49.00%	47.33%	54.33%	44.33%	47.88%**
	Seria A	1	47.30%	44.59%	48.30%	45.80%	46.64%	37.68%	45.05%**
		2	47.97%	44.30%	46.72%	45.30%	41.55%	42.28%	44.69%**

**Note:** The values in the table represent the OOS accuracy ratios. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner 68.75% is the OOS accuracy of CF for the 2009-2010 Premiership season and the 63.60% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship. A random classifier provides 33.33% accuracy ratio in this example. \*\* and \* indicates that according to the PT (1992) test, the forecasts are statistically accurate in classifying the football game result at the 99% and 95% level respectively.

Table 3.4: Accuracy Ratios (Asian Handicap)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	73.68%	78.57%	72.73%	61.11%	68.42%	66.67%	70.20%**
		2	56.25%	65.38%	66.67%	55.89%	63.16%	65.71%	62.18%**
	La -Liga	1	87.50%	69.62%	54.63%	66.67%	62.96%	58.33%	66.62%**
		2	63.60%	60.91%	62.71%	55.29%	64.28%	61.76%	61.43%**
	Seria A	1	71.88%	70.96%	85.71%	57.14%	70.58%	67.74%	70.67%**
		2	59.45%	62.07%	52.38%	53.06%	58.49%	51.43%	56.15%*
RVM	Premiership	1	69.07%	57.77%	54.36%	51.20%	54.00%	53.67%	56.68%
		2	58.45%	54.36%	47.77%	56.00%	50.67%	49.33%	52.76%
	La -Liga	1	72.67%	63.12%	55.82%	53.05%	55.67%	53.00%	58.89%*
		2	62.67%	56.51%	53.67%	57.33%	56.33%	55.33%	56.97%*
	Seria A	1	71.28%	59.12%	53.74%	49.16%	52.35%	50.70%	56.06%*
		2	56.76%	53.74%	50.00%	49.06%	51.06%	53.02%	52.27%
ANFIS	Premiership	1	54.00%	57.00%	52.92%	48.66%	49.32%	54.64%	52.76%
		2	52.70%	51.34%	47.08%	47.10%	50.33%	46.33%	49.15%
	La -Liga	1	54.00%	63.67%	57.88%	52.67%	52.00%	51.33%	55.26%
		2	54.67%	51.03%	52.67%	53.33%	50.10%	51.67%	52.25%
	Seria A	1	70.27%	57.77%	54.08%	52.52%	49.33%	49.29%	55.54%*
		2	57.77%	53.06%	44.54%	48.99%	50.35%	49.12%	50.64%
OP	Premiership	1	64.53%	58.45%	54.03%	46.05%	55.00%	48.67%	54.46%
		2	59.80%	54.36%	45.70%	50.00%	48.67%	47.33%	50.98%
	La -Liga	1	65.67%	57.67%	54.11%	53.00%	50.67%	49.00%	55.02%
		2	54.67%	55.14%	55.00%	49.67%	51.00%	49.00%	52.41%
	Seria A	1	66.22%	52.36%	52.72%	52.94%	47.32%	48.24%	53.30%
		2	56.42%	52.38%	52.94%	48.66%	48.59%	48.99%	51.33%

**Note:** The values in the table represent the OOS accuracy ratios. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 73.68% is the OOS accuracy of CF for the 2009-2010 Premiership season and the 56.25% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship. A random classifier provides 50.00% accuracy ratio in this example. \*\* and \* indicates that according to the PT (1992) test, the forecasts are statistically accurate in classifying the football game result at the 95% and 90% level respectively.

Table 3.5: Accuracy Ratios (Number of Goals)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	78.57%	73.68%	72.70%	62.50%	75.02%	69.23%	71.95%**
		2	75.71%	66.67%	71.43%	65.21%	70.59%	56.75%	67.73%**
	La -Liga	1	66.66%	65.93%	72.41%	72.02%	71.19%	62.50%	68.45%**
		2	60.87%	62.50%	62.61%	70.03%	72.70%	58.51%	64.54%**
	Seria A	1	85.69%	83.33%	70.00%	64.29%	71.43%	72.97%	74.62%**
		2	60.00%	52.17%	57.41%	58.62%	75.00%	59.46%	60.44%**
RVM	Premiership	1	50.17%	50.00%	55.37%	52.23%	51.67%	57.00%	52.74%**
		2	48.99%	55.03%	51.20%	51.67%	51.00%	53.00%	51.82%*
	La -Liga	1	57.00%	53.33%	53.42%	54.67%	51.00%	60.33%	54.96%**
		2	50.67%	52.05%	52.34%	53.30%	48.33%	54.67%	51.89%
	Seria A	1	53.04%	58.45%	53.74%	48.32%	48.32%	56.34%	53.04%*
		2	55.74%	51.36%	54.62%	47.65%	54.58%	50.34%	52.38%
ANFIS	Premiership	1	53.26%	50.14%	51.68%	47.08%	49.00%	53.00%	50.69%
		2	46.96%	50.34%	48.80%	55.33%	43.67%	53.67%	49.80%
	La -Liga	1	52.33%	52.00%	54.45%	50.33%	58.00%	49.33%	52.74%*
		2	52.00%	47.33%	53.67%	45.33%	52.05%	47.00%	49.56%
	Seria A	1	50.34%	53.04%	51.70%	52.94%	54.03%	61.62%	53.95%**
		2	47.30%	53.74%	44.96%	51.68%	56.34%	49.33%	50.56%
OP	Premiership	1	49.83%	49.66%	46.98%	49.48%	49.33%	55.33%	50.10%
		2	48.99%	52.68%	46.39%	51.33%	57.33%	51.00%	51.29%
	La -Liga	1	53.33%	54.67%	54.79%	52.67%	56.00%	56.10%	54.59%**
		2	48.00%	53.42%	53.35%	56.67%	57.00%	55.67%	54.02%**
	Seria A	1	46.62%	56.76%	50.68%	55.46%	47.32%	51.06%	51.32%
		2	45.13%	49.12%	47.28%	45.30%	50.70%	50.67%	48.03%

**Note:** The values in the table represent the OOS accuracy ratios. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 78.57% is the OOS accuracy of CF for the 2009-2010 Premiership season and the 75.71% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship. A random classifier provides 50.00% accuracy ratio in this example. \*\* and \* indicates that according to the PT (1992) test, the forecasts are statistically accurate in classifying the football game result at the 95% and 90% level respectively.

Table 3.6: Average Profit per Bet (Game Result)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	17.71%	19.33%	25.07%	26.81%	29.92%	14.50%	22.21%
		2	3.59%	10.00%	9.69%	0.52%	-0.50%	2.16%	4.24%
	La -Liga	1	10.89%	12.16%	10.45%	16.61%	7.92%	3.70%	10.29%
		2	1.13%	3.10%	2.15%	0.78%	2.80%	-1.09%	1.48%
	Seria A	1	5.17%	12.10%	23.19%	12.04%	7.24%	16.24%	12.66%
		2	-2.61%	-10.00%	1.42%	3.24%	3.54%	1.26%	-0.53%
RVM	Premiership	1	7.10%	-8.76%	-9.81%	1.09%	-1.80%	5.23%	-1.16%
		2	-6.00%	-13.74%	9.57%	-7.51%	-8.34%	-0.58%	-4.43%
	La -Liga	1	-5.45%	-0.45%	-0.32%	-1.05%	1.92%	2.03%	-0.55%
		2	-13.23%	-11.30%	-0.49%	-14.80%	-6.23%	-6.09%	-8.69%
	Seria A	1	4.58%	4.19%	2.79%	3.16%	-7.39%	-1.22%	1.02%
		2	-5.05%	-3.34%	-11.85%	-14.02%	-3.26%	-8.61%	-7.69%
ANFIS	Premiership	1	-2.57%	-3.07%	-2.91%	8.26%	-15.17%	-12.23%	-4.62%
		2	-7.09%	-12.56%	-21.83%	-5.06%	-5.43%	-7.09%	-9.84%
	La -Liga	1	-5.08%	-9.58%	-6.00%	-17.29%	-11.82%	0.40%	-8.23%
		2	-7.72%	-9.65%	-10.67%	-21.71%	-5.19%	-5.26%	-10.03%
	Seria A	1	-1.70%	-1.15%	2.63%	-1.34%	-3.22%	0.88%	-0.65%
		2	-18.26%	-7.33%	0.42%	-16.40%	-7.38%	-1.74%	-8.45%
OP	Premiership	1	-8.19%	1.24%	-8.07%	-16.31%	7.26%	-6.78%	-5.14%
		2	-9.02%	-15.86%	-17.71%	-8.05%	-7.31%	-7.29%	-10.88%
	La -Liga	1	-6.75%	-0.29%	7.50%	-2.65%	-25.15%	-0.62%	-4.66%
		2	-11.30%	-6.80%	-5.22%	-7.18%	1.30%	-13.29%	-7.08%
	Seria A	1	-3.76%	1.33%	-5.87%	-1.18%	-0.15%	-14.86%	-4.08%
		2	-4.42%	-1.02%	17.71%	-3.88%	-7.67%	-19.38%	-3.11%

**Note:** All values in the Table represent the average profit per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 17.71% is the average profit per bet of CF for the 2009-2010 Premiership season and the 3.59% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.7: Average Profit per Bet (Asian Handicap)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	25.00%	10.47%	12.53%	8.45%	26.09%	13.05%	15.93%
		2	2.65%	3.78%	6.02%	4.62%	-1.18%	-1.30%	2.43%
	La -Liga	1	24.08%	11.14%	9.36%	14.97%	15.20%	9.88%	14.11%
		2	2.63%	1.43%	2.14%	3.57%	1.98%	5.28%	2.84%
	Seria A	1	23.98%	13.73%	10.54%	7.79%	10.26%	16.58%	13.81%
		2	4.23%	-3.33%	2.52%	-4.14%	3.74%	3.45%	1.08%
RVM	Premiership	1	-0.58%	-4.63%	-0.37%	-6.20%	-2.96%	-4.07%	-3.14%
		2	-1.19%	-5.07%	-9.27%	-8.14%	-8.52%	-5.14%	-6.22%
	La -Liga	1	7.10%	3.02%	4.49%	-1.55%	-1.48%	1.62%	2.20%
		2	7.02%	1.30%	0.41%	-3.84%	-1.70%	1.32%	0.75%
	Seria A	1	4.04%	-2.18%	1.84%	-0.40%	-2.61%	-3.24%	-0.43%
		2	-4.03%	-0.15%	-5.54%	-3.42%	-13.06%	-7.14%	-5.56%
ANFIS	Premiership	1	-2.38%	0.04%	-8.72%	0.16%	0.51%	1.02%	-1.56%
		2	-5.24%	-11.74%	-10.26%	-4.44%	-5.13%	-1.42%	-6.37%
	La -Liga	1	-6.06%	8.77%	6.60%	0.64%	-2.73%	-5.22%	0.33%
		2	-6.40%	2.39%	-1.50%	-5.99%	-9.95%	0.26%	-3.53%
	Seria A	1	0.07%	-3.28%	2.89%	-4.25%	-6.73%	-2.32%	-2.27%
		2	-2.22%	-6.12%	-8.64%	-9.41%	-14.27%	-7.65%	-8.05%
OP	Premiership	1	2.97%	1.32%	0.90%	-8.38%	-0.82%	-5.79%	-1.63%
		2	0.03%	-0.14%	-8.02%	-7.19%	-8.18%	-11.03%	-5.76%
	La -Liga	1	-0.34%	-3.84%	-0.18%	-0.10%	-5.55%	0.50%	-1.59%
		2	-2.97%	0.45%	1.36%	-7.59%	-1.42%	-6.55%	-2.79%
	Seria A	1	6.45%	-5.97%	-1.04%	-0.01%	-5.22%	-10.61%	-2.73%
		2	0.56%	-4.96%	5.62%	-8.40%	-4.84%	-3.70%	-2.62%

**Note:** All values in the Table represent the average profit per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 25.00% is the average profit per bet of CF for the 2009-2010 Premiership season and the 2.65% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.8: Average Profit per Bet (Number of Goals)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	17.44%	16.25%	9.14%	11.64%	28.78%	17.62%	16.81%
		2	2.80%	-1.08%	-0.46%	2.83%	7.07%	4.77%	2.66%
	La -Liga	1	13.41%	15.44%	19.49%	14.14%	14.65%	9.87%	14.50%
		2	0.42%	2.15%	5.14%	0.28%	3.33%	0.83%	2.03%
	Seria A	1	17.22%	10.28%	6.45%	20.93%	14.32%	19.03%	14.71%
		2	1.20%	-0.96%	-0.25%	6.97%	2.44%	3.27%	2.11%
RVM	Premiership	1	1.77%	0.14%	-0.42%	-3.76%	-4.80%	-8.84%	-2.65%
		2	0.95%	-1.20%	-4.88%	-4.03%	-12.51%	-15.76%	-6.24%
	La -Liga	1	1.70%	-1.96%	-0.95%	-3.00%	4.11%	-3.06%	-0.53%
		2	-6.16%	-5.54%	-2.28%	-5.45%	-5.78%	-6.49%	-5.28%
	Seria A	1	-1.47%	-0.51%	-3.46%	-6.13%	-0.02%	-1.35%	-2.16%
		2	-1.53%	-2.79%	-6.56%	-9.97%	-1.58%	-6.40%	-4.81%
ANFIS	Premiership	1	-6.57%	-5.43%	0.12%	-3.22%	-5.20%	-3.38%	-3.95%
		2	-9.84%	-6.64%	-8.55%	-6.35%	-5.95%	-4.18%	-6.92%
	La -Liga	1	-4.20%	-2.12%	0.78%	-2.92%	0.12%	-3.66%	-2.00%
		2	-7.13%	-3.56%	-4.02%	-5.26%	-3.92%	-7.46%	-5.23%
	Seria A	1	-2.84%	-2.48%	-4.16%	-2.42%	-2.60%	-0.32%	-2.47%
		2	-4.61%	-5.07%	-10.49%	-6.93%	-3.30%	-5.02%	-5.90%
OP	Premiership	1	-7.95%	-11.28%	-0.16%	-3.86%	-1.96%	3.15%	-3.68%
		2	-7.80%	0.43%	-15.66%	-6.35%	5.78%	-9.50%	-5.52%
	La -Liga	1	-0.80%	-8.90%	-0.18%	-10.24%	-1.81%	-5.10%	-4.51%
		2	-8.75%	-5.94%	-10.17%	5.07%	-12.75%	-4.20%	-6.12%
	Seria A	1	-13.14%	-0.66%	-4.76%	0.03%	-13.08%	-10.74%	-7.06%
		2	4.59%	-9.44%	2.22%	-14.16%	-1.59%	-4.86%	-3.87%

**Note:** All values in the Table represent the average profit per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 17.44% is the average profit per bet of CF for the 2009-2010 Premiership season and the 2.80% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.9: Proportional Cumulative Annualized Return (Game Result)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	40.26%	7.42%	10.89%	11.22%	19.58%	6.91%	16.05%
		2	38.08%	1.35%	1.52%	-4.63%	11.11%	-5.70%	6.96%
	La -Liga	1	-0.14%	14.58%	22.39%	18.14%	29.46%	10.04%	15.75%
		2	-2.52%	3.25%	11.81%	-0.48%	-9.60%	-5.85%	-0.57%
	Seria A	1	15.34%	10.95%	28.64%	32.30%	32.01%	10.17%	21.57%
		2	-1.63%	-32.41%	2.01%	-5.02%	-17.08%	-6.29%	-10.07%
RVM	Premiership	1	-15.14%	-15.74%	-20.17%	-29.53%	-12.47%	-17.24%	-18.38%
		2	-27.12%	-45.78%	-54.28%	-35.52%	-60.05%	-58.30%	-46.84%
	La -Liga	1	-33.45%	-20.64%	-19.25%	-28.65%	-17.53%	-21.02%	-23.42%
		2	-55.26%	-57.36%	-38.27%	-47.20%	-42.08%	-47.36%	-47.92%
	Seria A	1	0.15%	-14.21%	-7.50%	-12.37%	-32.55%	-47.32%	-18.97%
		2	-25.13%	-60.93%	-16.06%	-61.27%	-59.77%	-62.83%	-47.67%
ANFIS	Premiership	1	-5.26%	-10.25%	-24.07%	-30.17%	-9.52%	-16.29%	-15.93%
		2	-29.16%	-15.57%	-36.49%	-33.60%	-23.41%	-23.47%	-26.95%
	La -Liga	1	-32.55%	-33.24%	-32.24%	-16.47%	-10.26%	-45.07%	-28.31%
		2	-68.26%	-38.56%	-69.41%	-23.54%	-15.05%	-52.14%	-44.49%
	Seria A	1	-55.85%	-26.37%	-6.84%	-47.29%	-14.54%	-13.25%	-27.36%
		2	-60.24%	-48.70%	-21.58%	-55.28%	-67.78%	-33.88%	-47.91%
OP	Premiership	1	-52.66%	-23.85%	-36.44%	-25.68%	-15.22%	-40.28%	-32.36%
		2	-64.66%	-32.10%	-52.13%	-43.69%	-19.65%	-56.09%	-44.72%
	La -Liga	1	-15.14%	-36.60%	-35.06%	-59.07%	-22.18%	-28.46%	-32.75%
		2	-18.44%	-68.12%	-62.38%	60.57%	-56.87%	-55.41%	-33.44%
	Seria A	1	-62.48%	-18.93%	-50.38%	-37.57%	-37.55%	-40.16%	-41.18%
		2	-66.64%	-46.18%	-55.37%	-66.88%	-59.85%	-58.74%	-58.94%

**Note:** All values in the Table represent the proportional cumulative annualized return per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 40.26% is the average profit per bet of CF for the 2009-2010 Premiership season and the 38.08% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.



Table 3.10: Proportional Cumulative Annualized Return (Asian Handicap)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	17.98%	13.23%	18.85%	12.57%	19.21%	16.93%	16.46%
		2	-9.83%	2.19%	17.00%	9.59%	-8.73%	-3.02%	1.20%
	La -Liga	1	22.20%	8.26%	6.05%	14.42%	18.11%	9.64%	13.11%
		2	3.25%	-0.12%	-5.47%	0.68%	0.12%	6.33%	0.80%
	Seria A	1	11.85%	6.50%	17.31%	6.19%	15.28%	12.94%	11.68%
		2	3.03%	1.30%	1.05%	-0.32%	9.27%	4.97%	3.22%
RVM	Premiership	1	-20.14%	-60.66%	-28.58%	-18.60%	-52.05%	-38.52%	-36.43%
		2	-33.96%	-64.59%	-32.55%	-68.63%	-59.15%	-59.33%	-53.04%
	La -Liga	1	-30.24%	-54.59%	-20.61%	-16.87%	-36.43%	-4.24%	-27.16%
		2	-44.39%	-58.12%	-47.02%	-43.38%	-41.82%	-9.25%	-40.66%
	Seria A	1	-39.25%	-31.86%	-40.58%	-53.99%	-50.13%	-32.31%	-41.35%
		2	-58.13%	-33.60%	-67.50%	-64.20%	-60.24%	-64.55%	-58.04%
ANFIS	Premiership	1	-12.47%	-26.63%	-20.55%	-26.30%	-19.34%	-13.76%	-19.84%
		2	-64.42%	-56.97%	-40.17%	48.59%	-65.24%	-39.75%	-36.33%
	La -Liga	1	-45.74%	-30.54%	-35.14%	-17.18%	-50.94%	-63.65%	-40.53%
		2	-54.80%	-60.04%	-43.11%	-32.82%	-55.87%	-65.71%	-52.06%
	Seria A	1	-33.75%	-40.40%	-50.14%	-50.79%	-41.72%	-20.20%	-39.50%
		2	-53.45%	-56.25%	-62.17%	-59.24%	-55.27%	-45.37%	-55.29%
OP	Premiership	1	-52.64%	-47.52%	-28.16%	-30.28%	-43.38%	-60.08%	-43.68%
		2	-64.03%	-61.47%	-51.21%	-55.02%	-65.17%	-68.67%	-60.93%
	La -Liga	1	-19.52%	-56.76%	-26.90%	-25.94%	-37.89%	-42.22%	-34.87%
		2	-50.73%	-61.23%	-56.47%	-49.57%	-46.37%	-57.03%	-53.57%
	Seria A	1	-54.07%	-42.87%	-34.31%	-50.37%	-61.68%	-48.39%	-48.62%
		2	-57.26%	-63.01%	-60.47%	-65.17%	-66.33%	-64.17%	-62.74%

**Note:** All values in the Table represent the proportional cumulative annualized return per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 17.98% is the average profit per bet of CF for the 2009-2010 Premiership season and the -9.83% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.11: Proportional Cumulative Annualized Return (Number of Goals)

Model	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
CF	Premiership	1	7.29%	24.58%	19.03%	6.68%	3.81%	33.46%	15.81%
		2	0.85%	8.04%	-1.69%	-3.42%	-2.03%	-6.21%	-0.74%
	La -Liga	1	10.24%	12.48%	5.87%	33.61%	9.56%	12.57%	14.06%
		2	-9.22%	4.65%	-8.56%	4.18%	2.55%	0.87%	-0.92%
	Seria A	1	9.17%	10.66%	11.54%	14.03%	25.19%	11.86%	13.74%
		2	3.63%	0.57%	3.12%	7.35%	-1.25%	2.63%	2.68%
RVM	Premiership	1	-21.14%	-15.58%	-31.61%	-58.47%	-14.87%	-18.40%	-26.68%
		2	-56.24%	-41.28%	-59.35%	-65.55%	-40.16%	-51.81%	-52.40%
	La -Liga	1	-26.24%	-16.38%	-46.72%	-57.43%	-32.55%	-30.04%	-34.89%
		2	-41.02%	-38.45%	-63.23%	-58.06%	-59.67%	-60.12%	-53.43%
	Seria A	1	-30.25%	-29.03%	-32.19%	-41.13%	-10.36%	-37.35%	-30.05%
		2	-65.02%	-50.14%	-48.21%	-56.01%	-40.62%	-46.18%	-51.03%
ANFIS	Premiership	1	-42.14%	-39.43%	-41.02%	-23.48%	-66.86%	-23.17%	-39.35%
		2	-49.56%	-51.17%	-49.52%	-49.08%	-70.35%	-49.61%	-53.22%
	La -Liga	1	-61.63%	-60.78%	-19.22%	-40.79%	-23.95%	-42.25%	-41.44%
		2	-63.28%	-61.36%	-53.24%	-67.59%	-59.31%	-52.22%	-59.50%
	Seria A	1	-52.78%	-55.90%	-56.32%	-50.63%	-57.05%	-54.00%	-54.45%
		2	-59.55%	-58.37%	-60.24%	-62.47%	-59.88%	-60.12%	-60.11%
OP	Premiership	1	-39.45%	-49.56%	-55.32%	-58.48%	-47.22%	-52.17%	-50.37%
		2	-60.17%	-58.16%	-59.42%	-62.21%	-55.66%	-63.14%	-59.79%
	La -Liga	1	-48.87%	-30.45%	-43.60%	-49.62%	-43.46%	-63.46%	-46.58%
		2	-65.24%	-48.80%	-66.75%	-63.96%	-60.17%	-62.85%	-61.30%
	Seria A	1	-41.13%	-32.87%	-63.55%	-60.32%	-57.08%	-58.45%	-52.23%
		2	-63.48%	-65.17%	-65.22%	-66.57%	-63.42%	-65.01%	-64.81%

**Note:** All values in the Table represent the proportional cumulative annualized return per season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 7.29% is the average profit per bet of CF for the 2009-2010 Premiership season and the 0.85% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.12: Kelly Criterion (CF)

			IS						
Exercise Championship	OOS		2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	Average
Game Result	Premiership	1	17.03%	20.27%	23.93%	25.90%	27.88%	15.34%	21.73%
		2	5.91%	12.45%	11.67%	4.88%	3.33%	6.16%	7.40%
	La -Liga	1	12.13%	15.27%	10.38%	17.57%	8.35%	4.91%	11.44%
		2	4.49%	7.66%	5.02%	6.32%	5.29%	3.45%	5.37%
	Seria A	1	6.17%	11.13%	24.61%	11.93%	8.10%	15.77%	12.95%
		2	2.78%	-4.50%	3.12%	8.12%	6.34%	6.25%	3.69%
Asian Handicap	Premiership	1	26.10%	10.43%	14.18%	7.01%	24.25%	13.33%	15.88%
		2	3.88%	5.51%	7.08%	4.90%	3.44%	3.67%	4.75%
	La -Liga	1	24.88%	11.52%	12.43%	15.90%	13.13%	10.88%	14.79%
		2	4.97%	3.11%	6.78%	5.69%	3.55%	7.28%	5.23%
	Seria A	1	21.23%	14.10%	12.15%	8.79%	9.31%	17.08%	13.78%
		2	5.22%	3.46%	4.57%	0.94%	4.74%	3.94%	3.81%
Number of Goals	Premiership	1	16.58%	17.31%	10.64%	12.23%	27.07%	19.84%	17.28%
		2	3.51%	2.37%	4.27%	4.59%	9.68%	6.54%	5.16%
	La -Liga	1	12.60%	15.86%	18.32%	15.91%	14.67%	11.32%	14.78%
		2	2.69%	3.23%	5.67%	4.25%	6.87%	3.85%	4.43%
	Seria A	1	17.54%	11.32%	6.03%	19.67%	15.41%	19.43%	14.90%
		2	5.20%	2.19%	3.14%	7.37%	6.42%	6.44%	5.13%

**Note:** All values in the Table represent the average profit per season of CF with the Kelly criterion. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 17.03% is the average profit per bet of CF for the 2009-2010 Premiership season and the 5.91% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 3.13: Kelly Criterion (OP)

Exercise	Championship	OOS	IS						Average
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	
Game Result	Premiership	1	0.53%	2.35%	1.59%	-3.08%	7.35%	-1.71%	1.17%
		2	-0.92%	-2.15%	-3.62%	-2.29%	-3.00%	-0.35%	-2.06%
	La -Liga	1	3.90%	0.41%	8.38%	-1.53%	-10.05%	-0.09%	0.17%
		2	1.45%	-4.17%	-3.84%	-5.16%	2.99%	-2.69%	-1.90%
	Seria A	1	-0.33%	2.61%	0.84%	0.36%	1.35%	-2.50%	0.39%
		2	-3.71%	-0.64%	-6.35%	-0.61%	-3.80%	-1.06%	-2.70%
	Premiership	1	4.77%	3.49%	1.62%	-3.25%	-0.08%	-1.14%	0.90%
		2	1.32%	0.41%	-3.65%	-2.09%	-3.43%	-1.35%	-1.47%
Asian Handicap	La -Liga	1	5.48%	-1.01%	11.45%	2.80%	-5.46%	-0.19%	2.18%
		2	3.90%	-2.34%	4.17%	-6.61%	0.08%	-0.90%	-0.28%
	Seria A	1	8.18%	-2.24%	1.05%	2.80%	-1.80%	-3.95%	0.67%
		2	1.01%	0.14%	-3.27%	-1.42%	-2.12%	-2.62%	-1.38%
	Premiership	1	-1.88%	-0.43%	0.78%	-1.03%	0.10%	5.48%	0.50%
		2	-0.76%	1.08%	-2.27%	0.39%	-1.27%	-1.68%	-0.75%
Number of Goals	La -Liga	1	-0.08%	1.38%	1.37%	-1.91%	-1.00%	-1.95%	-0.37%
		2	-0.61%	0.58%	-2.28%	-3.15%	-1.32%	-0.50%	-1.21%
	Seria A	1	4.26%	-0.08%	-2.58%	1.05%	-13.08%	6.58%	-0.64%
		2	-1.26%	-6.92%	-6.75%	-14.05%	-1.55%	-1.58%	-5.35%
	Premiership	1	-1.88%	-0.43%	0.78%	-1.03%	0.10%	5.48%	0.50%
		2	-0.76%	1.08%	-2.27%	0.39%	-1.27%	-1.68%	-0.75%

**Note:** All values in the Table represent the average profit per season of OP with the Kelly criterion. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 0.53% is the average profit per bet of CF for the 2009-2010 Premiership season and the -0.92% is the performance of the exact same model (same specification and rules) for the 2010-2011 season of the same championship.

Table 4.1: Summary statistics of the daily return series under study (12 MSCI indexes and the federal funds rate).

Market	Mean (bp)	Max (%)	Min (%)	Std. dev. (%)	Kurtosis	Skewness	First AC (significance)
<b>Developed</b>	1.55	9.10	-7.33	1.02	12.86	-0.50	0.12 (*)
US	1.45	11.04	-9.51	1.18	15.11	-0.36	-0.10 (*)
UK	1.60	17.32	-36.26	1.29	212.22	-6.62	0.01
Japan	2.66	12.77	-20.75	1.27	62.62	-2.12	-0.07
<b>Emerging</b>	1.91	10.07	-9.99	1.27	11.38	-0.49	0.22 (*)
Russia	1.83	42.37	-58.10	2.35	172.93	-2.26	0.02
China	3.32	14.05	-12.84	1.74	10.10	-0.04	0.03 (*)
Brazil	3.59	37.69	-46.23	2.19	109.54	-0.39	0.02
<b>Frontier</b>	1.95	12.54	-9.32	1.62	8.74	0.20	0.06 (*)
Estonia	2.96	5.50	-7.70	1.06	6.47	-0.13	0.16 (*)
Morocco	0.99	5.69	-9.07	0.83	15.56	-1.38	0.26 (*)
Jordan	1.21	7.82	-9.08	1.10	13.03	-0.71	0.07 (*)
<b>Federal funds rate</b>	0.53	0.02	0.00	0.01	2.78	1.18	1.00 (*)

**Note:** The mean daily returns are reported in basis points (bp). Maximum, minimum and standard deviation are presented in percentages (%). The last column reports the first-order autocorrelation coefficients. Coefficients notated with (\*) are significant at 1% (\*) level for the Ljung-Box Q statistic. The study period for all time-series is 01/01/2004 to 31/12/2016.

Table 4.2: Percentage and standard deviation of the DFDR<sup>±</sup> procedure survivors (IS 2 years).

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	0.34 (0.97)	1.16 (1.87)	0.43 (1.03)	2.39 (1.78)	2.84 (1.56)	3.52 (7.03)	0.24 (0.06)	3.50 (7.02)	16.24 (11.37)	0.33 (0.15)	<b>3.10</b> <b>(3.28)</b>
US	0.01 (0.00)	0.02 (0.01)	0.16 (0.32)	0.91 (1.12)	1.54 (0.94)	8.33 (9.49)	0.28 (0.07)	7.75 (10.55)	31.49 (3.55)	19.58 (13.81)	<b>7.01</b> <b>(3.99)</b>
UK	20.08 (9.08)	15.52 (10.26)	0.32 (0.38)	12.97 (10.21)	20.89 (6.96)	9.41 (9.88)	0.18 (0.04)	0.15 (0.06)	8.72 (11.71)	0.29 (0.09)	<b>8.85</b> <b>(5.87)</b>
Japan	3.59 (6.56)	0.11 (0.04)	0.69 (1.49)	3.20 (1.32)	2.58 (0.29)	1.09 (1.40)	0.29 (0.05)	0.27 (0.07)	0.18 (0.03)	0.16 (0.04)	<b>1.22</b> <b>(1.13)</b>
<b>Emerging</b>	2.87 (5.39)	1.26 (1.19)	1.12 (1.46)	3.35 (0.97)	6.04 (4.46)	9.01 (10.21)	0.36 (0.09)	0.39 (0.10)	0.19 (0.08)	0.24 (0.12)	<b>2.48</b> <b>(2.41)</b>
Russia	14.39 (7.39)	16.22 (9.55)	1.77 (3.48)	25.64 (7.84)	17.61 (4.78)	10.90 (10.65)	0.28 (0.10)	0.44 (0.11)	0.58 (0.14)	0.90 (0.30)	<b>8.87</b> <b>(4.43)</b>
China	2.59 (3.38)	32.32 (15.03)	5.14 (9.68)	0.93 (0.56)	3.72 (6.77)	3.59 (7.16)	0.28 (0.01)	0.27 (0.07)	0.20 (0.09)	0.81 (0.52)	<b>4.99</b> <b>(4.33)</b>
Brazil	22.52 (13.6)	8.62 (7.28)	8.84 (8.44)	17.79 (5.10)	20.84 (5.19)	8.71 (9.74)	0.09 (0.07)	0.27 (0.35)	1.02 (0.38)	0.32 (0.14)	<b>8.9</b> <b>(5.03)</b>
<b>Frontier</b>	14.14 (12.40)	1.10 (0.45)	3.85 (7.50)	29.22 (7.27)	25.26 (9.00)	7.24 (9.64)	0.37 (0.19)	0.47 (0.23)	4.39 (7.36)	1.49 (1.26)	<b>8.75</b> <b>(5.53)</b>
Estonia	17.37 (16.94)	0.35 (0.71)	4.34 (7.12)	7.91 (2.73)	7.52 (4.13)	10.23 (10.14)	0.19 (0.05)	4.35 (7.18)	8.86 (6.51)	0.66 (1.26)	<b>6.18</b> <b>(5.68)</b>
Morocco	7.36 (8.27)	26.96 (8.76)	17.24 (9.58)	4.82 (2.90)	0.64 (0.63)	0.15 (0.06)	0.34 (0.34)	0.65 (0.62)	0.15 (0.05)	0.22 (0.10)	<b>5.85</b> <b>(3.13)</b>
Jordan	20.26 (11.4)	1.57 (2.22)	1.77 (1.67)	4.27 (1.34)	1.52 (0.62)	0.67 (0.84)	0.21 (0.05)	0.09 (0.03)	0.11 (0.03)	0.18 (0.07)	<b>3.06</b> <b>(1.83)</b>
<b>Average</b>	<b>10.46</b> <b>(7.95)</b>	<b>8.77</b> <b>(4.78)</b>	<b>3.81</b> <b>(4.35)</b>	<b>9.45</b> <b>(3.60)</b>	<b>9.25</b> <b>(3.78)</b>	<b>6.07</b> <b>(7.19)</b>	<b>0.26</b> <b>(0.09)</b>	<b>1.55</b> <b>(2.20)</b>	<b>6.01</b> <b>(3.44)</b>	<b>2.10</b> <b>(1.49)</b>	<b>5.77</b> <b>(3.89)</b>

**Note:** The table reports the percentage and standard deviations of the survivor rules adjusted by the total number of rules. For example, in 2006 for the Developed market, the average number of surviving rules is 72 ( $0.0034 \times 21195$ ) and their standard deviation is 206 ( $0.0097 \times 21195$ ). The average is estimated from the twelve portfolios whose OOS is on 2006. The first portfolio's IS from 01/01/2004-31/12/2005 and the remaining eleven are calculated by rolling-forward the IS by one month.

Table 4.3: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	10.67%	11.87%	14.93%	27.00%	24.73%	8.65%	8.06%	7.92%	9.36%	5.64%	<b>12.88%</b>
	(1.62)	(1.44)	(1.39)	(1.66)	(1.62)	(1.79)	(1.75)	(1.65)	(1.57)	(2.39)	<b>(1.69)</b>
US	8.65%	10.41%	14.36%	25.44%	24.93%	9.54%	8.45%	10.73%	10.55%	7.50%	<b>13.05%</b>
	(0.95)	(1.11)	(1.11)	(1.19)	(1.20)	(1.38)	(1.50)	(1.51)	(1.48)	(1.24)	<b>(1.27)</b>
UK	10.13%	11.00%	15.50%	23.53%	28.08%	11.32%	4.89%	5.48%	11.53%	8.02%	<b>12.95%</b>
	(1.21)	(1.21)	(0.90)	(0.99)	(1.16)	(1.25)	(2.19)	(1.33)	(1.52)	(1.99)	<b>(1.38)</b>
Japan	14.23%	13.37%	9.13%	19.31%	18.57%	7.24%	7.05%	5.73%	3.70%	3.14%	<b>10.15%</b>
	(1.02)	(0.98)	(0.73)	(1.10)	(1.18)	(1.17)	(1.38)	(1.64)	(1.64)	(1.83)	<b>(1.27)</b>
<b>Emerging</b>	19.92%	22.49%	26.91%	37.28%	34.96%	12.36%	12.61%	12.86%	8.83%	6.25%	<b>19.45%</b>
	(2.35)	(2.09)	(1.84)	(2.04)	(1.98)	(1.88)	(2.08)	(2.09)	(2.11)	(1.89)	<b>(2.03)</b>
Russia	22.47%	22.45%	16.30%	46.59%	47.53%	15.61%	14.90%	17.32%	11.51%	22.86%	<b>23.76%</b>
	(0.95)	(0.98)	(0.70)	(1.18)	(1.30)	(1.15)	(1.74)	(1.75)	(1.76)	(1.65)	<b>(1.32)</b>
China	23.05%	24.14%	35.52%	43.71%	32.85%	8.73%	5.51%	8.40%	12.01%	14.51%	<b>20.84%</b>
	(1.65)	(1.73)	(1.68)	(1.52)	(1.37)	(1.37)	(2.33)	(1.90)	(1.75)	(1.65)	<b>(1.70)</b>
Brazil	24.90%	27.61%	30.54%	37.87%	35.75%	11.15%	11.41%	15.05%	13.20%	17.85%	<b>22.53%</b>
	(1.35)	(1.06)	(1.02)	(1.25)	(1.22)	(1.17)	(1.98)	(1.83)	(1.85)	(1.48)	<b>(1.42)</b>
<b>Frontier</b>	16.79%	17.46%	20.21%	29.42%	28.96%	12.98%	10.32%	10.23%	9.39%	10.49%	<b>16.62%</b>
	(2.60)	(2.33)	(1.95)	(2.06)	(2.24)	(2.39)	(2.17)	(2.24)	(2.27)	(2.02)	<b>(2.23)</b>
Estonia	18.07%	20.66%	29.00%	42.44%	40.75%	18.93%	13.62%	12.09%	13.37%	10.94%	<b>21.99%</b>
	(1.81)	(1.64)	(1.77)	(1.73)	(1.78)	(1.47)	(1.77)	(1.58)	(1.38)	(1.37)	<b>(1.63)</b>
Morocco	21.44%	21.55%	24.35%	27.58%	16.92%	4.60%	10.91%	12.83%	4.86%	3.47%	<b>14.85%</b>
	(2.07)	(1.91)	(1.71)	(1.60)	(1.54)	(1.99)	(1.17)	(1.21)	(1.95)	(1.34)	<b>(1.65)</b>
Jordan	33.77%	25.68%	21.99%	28.3%	22.09%	12.05%	9.5%	4.60%	6.92%	5.79%	<b>17.07%</b>
	(1.97)	(1.56)	(1.59)	(1.59)	(1.43)	(1.79)	(1.83)	(1.91)	(1.52)	(1.55)	<b>(1.67)</b>
<b>Average</b>	<b>18.67%</b>	<b>19.06%</b>	<b>21.56%</b>	<b>32.37%</b>	<b>29.68%</b>	<b>11.10%</b>	<b>9.77%</b>	<b>10.27%</b>	<b>9.60%</b>	<b>9.70%</b>	<b>17.18%</b>
	<b>(1.63)</b>	<b>(1.50)</b>	<b>(1.37)</b>	<b>(1.49)</b>	<b>(1.50)</b>	<b>(1.57)</b>	<b>(1.82)</b>	<b>(1.72)</b>	<b>(1.73)</b>	<b>(1.70)</b>	<b>(1.60)</b>

**Note:** The table reports the average IS annualized returns and Sharpe ratios of twelve portfolios for two years of IS after transaction costs (rolling-forward by one month). For example, the 10.67% annualized return of the Developed markets (2006) is calculated as the average IS annualized return of twelve portfolios. The first portfolio's IS return is calculated over the period of 01/01/2004-31/12/2005. The remaining eleven are calculated by rolling-forward the IS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table 4.4: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 1 Month)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	-3.66%	-4.72%	10.63%	4.14%	-17.79%	-1.92%	-3.23%	0.35%	-2.07%	-1.43%	<b>-1.97%</b>
	(-0.54)	(-0.43)	(0.41)	(0.25)	(-2.91)	(-0.5)	(-1.27)	(0.13)	(-0.42)	(-0.64)	<b>(-0.59)</b>
US	1.81%	-6.16%	22.72%	-3.64%	-14.07%	-1.65%	-1.37%	7.02%	5.73%	-5.28%	<b>0.51%</b>
	(0.2)	(-0.45)	(0.77)	(-0.19)	(-1.49)	(-0.31)	(-0.5)	(1.31)	(0.77)	(-0.71)	<b>(-0.06)</b>
UK	15.94%	-3.14%	16.5%	19.58%	-12.73%	-4.12%	-3.83%	1.67%	-7.09%	-2.8%	<b>2%</b>
	(1.55)	(-0.27)	(0.41)	(1.16)	(-1.47)	(-0.58)	(-3.7)	(0.67)	(-0.97)	(-0.45)	<b>(-0.36)</b>
Japan	-11.33%	-8.68%	41.53%	11.47%	-2.11%	-7.87%	-3.3%	-3.95%	-5.75%	-2.98%	<b>0.7%</b>
	(-0.76)	(-1.8)	(1.4)	(0.51)	(-0.37)	(-1.62)	(-1.31)	(-1.03)	(-2.27)	(-1.13)	<b>(-0.84)</b>
<b>Emerging</b>	1.75%	-1.93%	43.52%	4.25%	-7.22%	-3.14%	-0.93%	-1.71%	-3.71%	-1.16%	<b>2.97%</b>
	(0.15)	(-0.15)	(1.1)	(0.29)	(-1.09)	(-0.52)	(-0.23)	(-0.55)	(-1.28)	(-0.31)	<b>(-0.26)</b>
Russia	45.77%	-9%	45.64%	13.09%	-20.64%	-9.9%	-2.42%	-4.63%	6.92%	-12.48%	<b>5.24%</b>
	(1.25)	(-1.33)	(1.01)	(0.47)	(-2.1)	(-1.02)	(-0.44)	(-0.92)	(0.44)	(-0.98)	<b>(-0.36)</b>
China	59.59%	29.16%	-2.1%	-19.48%	-8.93%	-10.76%	-0.29%	-3.54%	-4.51%	11.46%	<b>5.06%</b>
	(2.62)	(1.17)	(-0.05)	(-1.04)	(-1.09)	(-1.4)	(-0.64)	(-1.42)	(-0.66)	(0.82)	<b>(-0.17)</b>
Brazil	7.53%	67.85%	48.86%	15.4%	-15.81%	-5.12%	-7.98%	-3.03%	-7.35%	-1.07%	<b>9.93%</b>
	(0.28)	(1.2)	(0.9)	(0.68)	(-1.67)	(-0.73)	(-2.42)	(-0.49)	(-0.6)	(-0.07)	<b>(-0.29)</b>
<b>Frontier</b>	-11.64%	14.24%	64.67%	11.12%	0.26%	-12.4%	-0.27%	4.2%	7.75%	2.33%	<b>8.03%</b>
	(-2.04)	(1.44)	(2.24)	(1.04)	(0.06)	(-3.55)	(-0.09)	(0.98)	(1.39)	(0.32)	<b>(0.18)</b>
Estonia	-4.93%	-7.76%	65.08%	2.83%	6.06%	-24.12%	6.26%	-4.29%	12.92%	-13.23%	<b>3.88%</b>
	(-0.62)	(-0.45)	(1.46)	(0.1)	(0.3)	(-2.6)	(0.98)	(-0.69)	(1.23)	(-2.11)	<b>(-0.24)</b>
Morocco	34%	21.97%	25.32%	1.5%	-10.22%	-2.49%	1.19%	-4.9%	-0.64%	-0.36%	<b>6.54%</b>
	(1.87)	(1.65)	(1.2)	(0.1)	(-2.17)	(-0.87)	(0.1)	(-0.48)	(-0.61)	(-0.17)	<b>(0.06)</b>
Jordan	-0.62%	-3.47%	37.89%	-9.46%	-0.12%	-0.79%	-4.59%	-2.27%	-4.63%	-1.32%	<b>1.06%</b>
	(-0.04)	(-0.43)	(1.32)	(-0.81)	(-0.02)	(-0.13)	(-1.55)	(-1.03)	(-1.23)	(-0.37)	<b>(-0.43)</b>
<b>Average</b>	11.18%	7.36%	35.02%	4.23%	-8.61%	-7.02%	-1.73%	-1.26%	-0.2%	-2.36%	<b>3.66%</b>
	(0.33)	(0.01)	(1.02)	(0.21)	(-1.17)	(-1.15)	(-0.92)	(-0.29)	(-0.35)	(-0.48)	<b>(-0.28)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of twelve portfolios for two years of IS and one month of OOS after transaction costs (rolling-forward by one month). For example, the -3.66% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over January 2006 using as IS the period 01/01/2004-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.



Table 4.5: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 3 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	-4.54% (-0.62)	-3.26% (-0.27)	24.85% (0.99)	-5.42% (-0.4)	-13.37% (-2.23)	-2.42% (-0.53)	-2.91% (-1.22)	-1.35% (-0.46)	0.20% (0.04)	-2.24% (-1.06)	<b>-1.05%</b> <b>(-0.57)</b>
US	4.28% (0.48)	-4.93% (-0.36)	34.43% (1.12)	-14.00% (-0.89)	-9.63% (-1.10)	-5.86% (-0.86)	-0.98% (-0.41)	5.41% (1.03)	6.43% (0.86)	-4.84% (-0.68)	<b>1.03%</b> <b>(-0.08)</b>
UK	14.9% (1.48)	-13.25% (-0.88)	35.36% (0.79)	9.70% (0.67)	-9.29% (-1.03)	-5.68% (-0.8)	-2.07% (-2.10)	0.82% (0.31)	-5.53% (-0.81)	-2.75% (-0.46)	<b>2.22%</b> <b>(-0.28)</b>
Japan	-16.79% (-0.96)	-9.00% (-1.81)	29.70% (1.03)	4.98% (0.23)	-1.76% (-0.35)	-5.78% (-1.53)	-3.03% (-1.25)	-2.77% (-0.86)	-2.84% (-1.70)	-3.67% (-1.49)	<b>-1.10%</b> <b>(-0.87)</b>
<b>Emerging</b>	-4.31% (-0.42)	-2.01% (-0.15)	15.33% (0.49)	-1.19% (-0.10)	-5.19% (-0.79)	-0.19% (-0.03)	-1.92% (-0.52)	-0.38% (-0.11)	-4.35% (-1.74)	-1.54% (-0.4)	<b>-0.57%</b> <b>(-0.38)</b>
Russia	37.24% (1.11)	-6.90% (-0.54)	31.94% (0.95)	3.01% (0.13)	-17.55% (-1.79)	-7.51% (-0.82)	-0.09% (-0.02)	-5.58% (-1.08)	0.52% (0.04)	-12.47% (-1.08)	<b>2.26%</b> <b>(-0.31)</b>
China	35.48% (1.70)	29.13% (1.20)	3.30% (0.09)	-8.37% (-0.51)	-6.97% (-0.89)	-4.61% (-0.71)	-0.42% (-0.87)	-3.18% (-1.24)	-6.54% (-1.04)	3.89% (0.32)	<b>4.17%</b> <b>(-0.19)</b>
Brazil	1.52% (0.06)	50.70% (1.05)	7.94% (0.18)	11.17% (0.52)	-14.63% (-1.55)	-3.96% (-0.59)	-5.03% (-1.62)	-5.62% (-0.91)	-7.23% (-0.65)	-2.33% (-0.18)	<b>3.25%</b> <b>(-0.37)</b>
<b>Frontier</b>	-11.37% (-2.03)	9.60% (1.01)	75.78% (2.65)	-2.44% (-0.31)	2.13% (0.50)	-6.28% (-1.67)	0.37% (0.13)	1.03% (0.26)	1.24% (0.25)	4.47% (0.60)	<b>7.45%</b> <b>(0.14)</b>
Estonia	-1.93% (-0.18)	-11.44% (-0.79)	55.63% (1.37)	11.29% (0.40)	-3.33% (-0.20)	-17.59% (-1.95)	0.50% (0.09)	-4.46% (-0.83)	5.34% (0.53)	-8.70% (-1.61)	<b>2.53%</b> <b>(-0.32)</b>
Morocco	17.57% (1.12)	24.57% (1.92)	9.62% (0.55)	-7.56% (-0.63)	-8.20% (-1.83)	-2.49% (-0.87)	-1.49% (-0.15)	-6.25% (-0.62)	-0.98% (-0.83)	-2.83% (-1.49)	<b>2.19%</b> <b>(-0.28)</b>
Jordan	-2.77% (-0.20)	-1.36% (-0.16)	32.59% (1.22)	-12.14% (-1.12)	-0.37% (-0.05)	-0.28% (-0.05)	-4.27% (-1.46)	-2.94% (-1.19)	-2.37% (-0.65)	-0.49% (-0.14)	<b>0.56%</b> <b>(-0.38)</b>
<b>Average</b>	<b>5.77%</b> <b>(0.13)</b>	<b>5.15%</b> <b>(0.02)</b>	<b>29.71%</b> <b>(0.95)</b>	<b>-0.91%</b> <b>(-0.17)</b>	<b>-7.35%</b> <b>(-0.94)</b>	<b>-5.22%</b> <b>(-0.87)</b>	<b>-1.78%</b> <b>(-0.78)</b>	<b>-2.11%</b> <b>(-0.47)</b>	<b>-1.34%</b> <b>(-0.48)</b>	<b>-2.79%</b> <b>(-0.64)</b>	<b>1.91%</b> <b>(-0.33)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of four portfolios for IS of two years and OOS of three months after transaction costs (rolling-forward by one month). For example, the -4.54% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over the period 01/01/2006-31/03/2006 using as IS the period 01/01/2004-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS and the OOS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table 4.6: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 2 Years and OOS 6 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	1.05% (0.14)	-1.88% (-0.15)	13.92% (0.57)	-2.48% (-0.22)	-9.35% (-1.49)	-5.75% (-1.16)	-2.58% (-1.14)	0.74% (0.26)	-0.82% (-0.17)	-0.75% (-0.32)	<b>-0.79%</b> <b>(-0.37)</b>
US	0.53% (0.06)	-6.82% (-0.46)	26.57% (0.90)	-7.22% (-0.53)	-6.95% (-0.82)	-5.85% (-0.72)	0.63% (0.25)	4.77% (0.95)	5.66% (0.75)	-6.46% (-0.95)	<b>0.49%</b> <b>(-0.06)</b>
UK	15.09% (1.45)	-12.07% (-0.91)	42.99% (0.98)	1.97% (0.16)	-8.36% (-0.92)	-7.78% (-0.93)	-1.18% (-1.27)	0.68% (0.29)	-7.72% (-1.20)	-2.01% (-0.32)	<b>2.16%</b> <b>(-0.27)</b>
Japan	-18.42% (-1.02)	-9.13% (-1.59)	35.99% (1.11)	0.61% (0.04)	-2.46% (-0.51)	-3.60% (-1.18)	-2.15% (-0.83)	-2.42% (-0.88)	-1.25% (-0.86)	-3.34% (-1.29)	<b>-0.62%</b> <b>(-0.70)</b>
<b>Emerging</b>	-1.43% (-0.15)	4.61% (0.34)	18.10% (0.57)	0.35% (0.03)	-2.69% (-0.41)	-4.97% (-0.78)	0.87% (0.21)	-0.65% (-0.18)	-2.32% (-0.88)	-1.87% (-0.45)	<b>1.00%</b> <b>(-0.17)</b>
Russia	17.28% (0.68)	-2.67% (-0.20)	34.27% (0.94)	-4.48% (-0.22)	-15.42% (-1.54)	-11.21% (-1.05)	-1.32% (-0.28)	-10.12% (-1.83)	0.31% (0.02)	-5.83% (-0.48)	<b>0.08%</b> <b>(-0.39)</b>
China	25.95% (1.35)	25.27% (1.01)	0.34% (0.01)	-5.22% (-0.33)	-3.90% (-0.48)	-5.51% (-0.76)	-0.45% (-0.91)	-4.18% (-1.71)	-5.06% (-0.78)	0.65% (0.06)	<b>2.79%</b> <b>(-0.25)</b>
Brazil	-5.00% (-0.22)	42.45% (0.99)	-24.49% (-0.66)	11.68% (0.57)	-14.25% (-1.54)	-8.22% (-1.00)	-3.92% (-1.24)	-2.79% (-0.42)	-11.07% (-0.98)	5.69% (0.40)	<b>-0.99%</b> <b>(-0.41)</b>
<b>Frontier</b>	-7.46% (-1.31)	3.34% (0.37)	59.03% (2.54)	-2.60% (-0.39)	1.46% (0.34)	-6.31% (-1.78)	1.02% (0.36)	1.84% (0.48)	-3.06% (-0.68)	5.85% (0.79)	<b>5.31%</b> <b>(0.07)</b>
Estonia	-2.18% (-0.18)	0.50% (0.04)	43.59% (1.21)	13.47% (0.50)	-4.13% (-0.29)	-17.81% (-1.96)	3.27% (0.58)	-5.76% (-1.20)	0.73% (0.08)	-4.64% (-0.74)	<b>2.7%</b> <b>(-0.2)</b>
Morocco	11.92% (0.83)	20.50% (1.67)	-3.46% (-0.25)	-10.05% (-0.92)	-7.16% (-1.46)	-2.89% (-1.13)	-2.86% (-0.30)	-5.93% (-0.62)	-1.86% (-1.42)	-2.17% (-1.07)	<b>-0.4%</b> <b>(-0.47)</b>
Jordan	-6.88% (-0.57)	-0.36% (-0.04)	29.07% (1.17)	-12.65% (-1.22)	0.07% (0.01)	-3.22% (-0.61)	-4.53% (-1.75)	-2.88% (-1.14)	-2.6% (-0.76)	-0.29% (-0.09)	<b>-0.43%</b> <b>(-0.50)</b>
<b>Average</b>	<b>2.54%</b> <b>(0.09)</b>	<b>5.31%</b> <b>(0.09)</b>	<b>22.99%</b> <b>(0.76)</b>	<b>-1.39%</b> <b>(-0.21)</b>	<b>-6.09%</b> <b>(-0.76)</b>	<b>-6.93%</b> <b>(-1.09)</b>	<b>-1.10%</b> <b>(-0.53)</b>	<b>-2.23%</b> <b>(-0.50)</b>	<b>-2.42%</b> <b>(-0.57)</b>	<b>-1.27%</b> <b>(-0.37)</b>	<b>0.94%</b> <b>(-0.31)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of two portfolios for IS of two years and OOS of six months after transaction costs (rolling-forward by one month). For example, the 1.05% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over the period 01/01/2006-30/06/2006 using as IS the period 01/01/2004-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS and the OOS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table 4.7: Monthly Performance Persistence for IS 2 Years (1 month rolling-forward)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	1.00	0.58	0.75	0.83	0.42	0.75	0.50	0.67	0.83	0.75	<b>0.71</b>
US	1.17	0.42	1.08	1.08	1.33	1.17	1.00	1.92	0.83	0.25	<b>1.03</b>
UK	0.92	0.58	0.50	1.08	0.42	0.58	0.08	0.75	0.42	1.00	<b>0.63</b>
Japan	0.17	0.50	0.58	0.92	1.58	0.58	0.25	0.17	0.08	0.42	<b>0.53</b>
<b>Emerging</b>	0.50	0.92	1.17	1.33	0.33	0.75	0.58	0.17	0.17	0.83	<b>0.68</b>
Russia	0.42	0.25	0.17	0.33	0.33	0.67	0.67	0.33	0.92	0.33	<b>0.44</b>
China	2.17	2.92	0.42	0.58	0.42	0.42	0.83	0.42	0.67	1.08	<b>0.99</b>
Brazil	0.67	0.58	0.67	0.75	0.33	0.42	0.33	0.58	0.92	0.50	<b>0.58</b>
<b>Frontier</b>	0.42	1.67	2.25	0.50	0.58	0.17	0.75	0.67	1.33	1.25	<b>0.96</b>
Estonia	0.50	0.42	0.75	0.58	0.75	0.17	0.67	0.42	1.08	0.25	<b>0.56</b>
Morocco	1.92	2.00	1.25	0.42	0.25	0.33	0.58	0.92	0.42	0.58	<b>0.87</b>
Jordan	0.42	0.75	2.08	0.50	0.83	0.83	0.25	0.75	0.58	0.58	<b>0.76</b>
<b>Average</b>	<b>0.85</b>	<b>0.97</b>	<b>0.97</b>	<b>0.74</b>	<b>0.63</b>	<b>0.57</b>	<b>0.54</b>	<b>0.65</b>	<b>0.69</b>	<b>0.65</b>	<b>0.73</b>

**Note:** The table reports the average number of consecutive months that the monthly OOS returns of the twelve portfolio returns are above the risk-free rate. This average is calculated by generating the monthly OOS in consecutive months for each of the twelve portfolios mentioned in Table 4.4. For example, in Developed markets (2006) for the first portfolio, I calculate the OOS returns for 2006 (January, February, etc.). If the OOS returns over the first month are below the relevant risk-free rate, I assign a value of 0. If the OOS returns remain above the risk-free rate during the first month e.g. in January but not for February, I assign the value of 1. Otherwise, I assign a value of 2 or more. This process is repeated for the remaining eleven portfolios of the year. The analysis is done using the maximum 18 months of OOS calculations for each portfolio. The last column and row presents the average monthly performance persistence per market across all years and per year respectively.

Table 4.8: Quarterly Performance Persistence in Months for IS 2 2 Years (3 months rolling-forward)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	0.33	0.67	1.08	0.75	0.42	0.33	0.25	0.58	1.17	0.50	<b>0.61</b>
US	1.17	1.00	2.58	0.42	0.58	0.50	0.58	1.67	2.17	0.42	<b>1.11</b>
UK	2.17	0.58	1.17	0.67	0.33	0.33	0.00	0.75	0.42	0.50	<b>0.69</b>
Japan	0.33	0.17	1.33	0.83	0.67	0.17	0.17	0.33	0.08	0.25	<b>0.43</b>
<b>Emerging</b>	0.83	0.67	0.83	1.08	0.33	0.67	1.42	0.50	0.17	0.75	<b>0.73</b>
Russia	1.50	0.50	0.83	0.67	0.33	0.33	0.92	0.25	0.83	0.33	<b>0.65</b>
China	1.75	1.08	0.42	0.17	0.42	0.42	0.42	0.08	0.58	0.75	<b>0.61</b>
Brazil	1.00	0.58	0.50	1.08	0.17	0.25	0.33	0.58	0.75	1.00	<b>0.63</b>
<b>Frontier</b>	0.50	1.42	1.25	0.50	0.75	0.25	1.33	0.75	0.42	1.25	<b>0.84</b>
Estonia	0.25	0.67	1.25	0.75	0.67	0.08	0.75	0.25	1.08	0.25	<b>0.60</b>
Morocco	1.00	2.83	0.67	0.33	0.00	0.17	0.33	0.42	0.25	0.17	<b>0.62</b>
Jordan	0.42	0.75	1.17	0.00	0.42	0.58	0.42	0.17	0.67	0.50	<b>0.51</b>
<b>Average</b>	<b>0.94</b>	<b>0.91</b>	<b>1.09</b>	<b>0.60</b>	<b>0.42</b>	<b>0.34</b>	<b>0.58</b>	<b>0.53</b>	<b>0.72</b>	<b>0.56</b>	<b>0.67</b>

**Note:** The table reports the average number of consecutive months that the quarterly OOS returns of the twelve portfolio returns are above the risk-free rate. This average is calculated by generating the quarterly OOS in consecutive quarters for each of the twelve portfolios mentioned in Table 4.5. For example, in Developed markets (2006) for the first portfolio, I calculate the OOS returns for 2006 (January to March, February to April, etc.). If the OOS returns over the first three months are below the relevant risk-free rate, I assign a value of 0. If the OOS returns remain above the risk-free rate only during the first 3 months of the OOS e.g. in January to March but not from February to June, I assign the value of 1. Otherwise, I assign a value of 2 or more. This process is repeated for the remaining eleven portfolios of the year. The analysis is done using a maximum of 18 months (6 quarters) of OOS calculations for each portfolio. The last column and row presents the average performance per market across all years and per year respectively.

Table 4.9: Semi-annual Performance Persistence in Months for IS 2 Years (6 months rolling-forward)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	0.50	0.42	1.25	0.50	0.42	0.08	0.17	0.75	0.58	0.33	<b>0.50</b>
US	1.00	0.75	1.67	0.33	0.42	0.17	1.17	2.00	1.83	0.17	<b>0.95</b>
UK	1.83	0.17	0.92	0.67	0.08	0.00	0.00	0.42	0.25	0.50	<b>0.48</b>
Japan	0.00	0.33	1.33	0.50	0.25	0.25	0.17	0.17	0.08	0.08	<b>0.32</b>
<b>Emerging</b>	1.25	0.67	1.42	1.00	0.50	0.42	1.25	0.42	0.67	0.25	<b>0.78</b>
Russia	1.25	0.42	1.17	0.50	0.08	0.33	0.67	0.00	0.67	0.17	<b>0.53</b>
China	1.58	0.83	1.17	0.42	0.50	0.08	0.00	0.00	0.08	0.83	<b>0.55</b>
Brazil	1.33	1.83	0.67	1.00	0.08	0.00	0.08	0.25	0.25	1.25	<b>0.68</b>
<b>Frontier</b>	0.67	0.75	1.50	0.33	0.50	0.08	1.08	1.33	0.75	1.00	<b>0.80</b>
Estonia	0.50	0.75	2.08	1.25	0.50	0.00	0.67	0.00	0.67	0.42	<b>0.68</b>
Morocco	1.58	1.83	0.42	0.00	0.00	0.08	0.50	0.17	0.00	0.08	<b>0.47</b>
Jordan	0.33	1.00	1.83	0.00	0.67	0.25	0.00	0.00	0.75	1.08	<b>0.59</b>
<b>Average</b>	<b>0.99</b>	<b>0.81</b>	<b>1.28</b>	<b>0.54</b>	<b>0.33</b>	<b>0.15</b>	<b>0.48</b>	<b>0.46</b>	<b>0.55</b>	<b>0.51</b>	<b>0.61</b>

**Note:** The table reports the average number of consecutive months that the semi-annual OOS returns of the twelve portfolio returns are above the risk-free rate. This average is calculated by generating the semi-annual OOS in consecutive quarters for each of the twelve portfolios mentioned in Table 4.6. For example, in Developed markets (2006) for the first portfolio, I calculate the OOS returns for 2006 (January to June, February to July, etc.). If the OOS returns over the first six months are below the relevant risk-free rate, I assign a value of 0. If the OOS returns remain above the risk-free rate only during the first 6 months of the OOS e.g. in January to June but not from July to December, I assign the value of 1. Otherwise, I assign a value of 2 or more. This process is repeated for the remaining eleven portfolios of the year. The analysis is done using a maximum of 18 months (3 semesters) of OOS calculations for each portfolio. The last column and row presents the average performance per market across all years and per year respectively.

Table 4.10: Annualized Returns Based on the Cross-validated Surviving Rules (IS of 2 Years and OOS 1 Month)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	27.05% (0.02%)	23.08% (0.38%)	103.68% (0.09%)	93.3% (1%)	45.17% (0.73%)	46.44% (3.19%)	36.01% (0.03%)	17.92% (3.22%)	21.53% (7.76%)	27.87% (0.06%)	<b>44.21%</b> <b>(1.65%)</b>
US	8.19% (0%)	18.04% (0.01%)	87.67% (0.04%)	84.56% (0.33%)	50.42% (0.38%)	34.72% (4.82%)	24.13% (0.05%)	26.95% (7.41%)	21.08% (17.41%)	21.33% (8.32%)	<b>37.71%</b> <b>(3.88%)</b>
UK	23.8% (9.29%)	27.1% (6.91%)	89.14% (0.08%)	85.49% (6.76%)	56.82% (7.79%)	50.13% (5.17%)	27.01% (0.01%)	20.6% (0.03%)	40.89% (4.1%)	36.97% (0.07%)	<b>45.8%</b> <b>(4.02%)</b>
Japan	29.93% (0.02%)	6.87% (0.02%)	59.93% (0.02%)	66.47% (1.3%)	39.6% (0.61%)	32.77% (0.14%)	36.79% (0.02%)	29.92% (0.03%)	6.11% (0.01%)	9.93% (0.01%)	<b>31.83%</b> <b>(0.22%)</b>
<b>Emerging</b>	58.41% (0.58%)	70.48% (0.57%)	163.42% (0.54%)	116.68% (1.7%)	51.44% (2.99%)	53.04% (1.74%)	39.71% (0.08%)	27.96% (0.09%)	30.23% (0.05%)	47.04% (0.04%)	<b>65.84%</b> <b>(0.84%)</b>
Russia	86.02% (4.44%)	27.64% (2.67%)	140.78% (0.85%)	183.65% (8.59%)	67.8% (6.09%)	74.31% (6.62%)	75.11% (0.06%)	44.42% (0.11%)	109.66% (0.13%)	91.44% (0.23%)	<b>90.08%</b> <b>(2.98%)</b>
China	99.78% (2.08%)	99.23% (18.07%)	170% (1.23%)	104.07% (0.35%)	55.13% (0.18%)	49.38% (1.62%)	43.64% (0.01%)	27.92% (0.03%)	37.37% (0.07%)	78.59% (0.21%)	<b>76.51%</b> <b>(2.38%)</b>
Brazil	106.43% (6.81%)	103.96% (3.32%)	172.44% (1.62%)	122.07% (6.6%)	65.04% (7.76%)	39.45% (3.15%)	36.28% (0.01%)	51.21% (0.08%)	59.73% (0.31%)	79.44% (0.08%)	<b>83.6%</b> <b>(2.97%)</b>
<b>Frontier</b>	34.04% (5.02%)	45.26% (0.65%)	131.42% (2.91%)	90.36% (14.16%)	40.02% (10.53%)	30.28% (2.11%)	26.5% (0.12%)	25.28% (0.18%)	26.89% (2.37%)	35.09% (0.71%)	<b>48.51%</b> <b>(3.88%)</b>
Estonia	42.02% (7.72%)	64.11% (0.1%)	246.5% (2.06%)	119.29% (3.52%)	141.52% (3.22%)	52.77% (1.96%)	43.17% (0.04%)	31.11% (1.68%)	51.85% (2.55%)	29.85% (0.03%)	<b>82.22%</b> <b>(2.29%)</b>
Morocco	97.25% (4.06%)	42.05% (18.14%)	77.99% (9.32%)	51.52% (1.75%)	37.84% (0.11%)	16% (0.02%)	36.57% (0.07%)	40.51% (0.29%)	10.15% (0.01%)	16.5% (0.03%)	<b>42.64%</b> <b>(3.38%)</b>
Jordan	89.39% (7.42%)	45.23% (0.46%)	103.32% (0.93%)	52.78% (1.82%)	35.76% (0.59%)	37.43% (0.24%)	27.1% (0.05%)	14.9% (0.01%)	18.51% (0.03%)	22.1% (0.03%)	<b>44.65%</b> <b>(1.16%)</b>
<b>Average</b>	<b>58.52%</b> <b>(3.95%)</b>	<b>47.75%</b> <b>(4.27%)</b>	<b>128.86%</b> <b>(1.64%)</b>	<b>97.52%</b> <b>(3.99%)</b>	<b>57.21%</b> <b>(3.41%)</b>	<b>43.06%</b> <b>(2.56%)</b>	<b>37.67%</b> <b>(0.05%)</b>	<b>29.89%</b> <b>(1.1%)</b>	<b>36.17%</b> <b>(2.9%)</b>	<b>41.35%</b> <b>(0.82%)</b>	<b>57.80%</b> <b>(2.47%)</b>

**Note:** The table reports the average OOS annualized returns of twelve cross-validated portfolios for IS of two years and OOS of one month after transaction costs (rolling-forward by one month). In the parentheses, I report the average percentage of cross-validated rules from the total pool of rules. For example, Table 2 reports that 2628 rules ( $0.124 \times 21195$ ) survive on average in the case of frontier markets (2006). This table estimates that from those rules, 132 rules ( $0.0502 \times 2628$ ) are surviving both in the IS and OOS and they generate an average OOS annualized return of 34.04%.

Table 4.11: Annualized Returns Based on the Cross-validated Surviving Rules (IS of 2 Years and OOS 3 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	12.96% (0.02%)	15.73% (0.04%)	64.58% (0.3%)	36.66% (0.85%)	21.51% (0.6%)	22.21% (0.09%)	20.42% (0.04%)	12.8% (3.17%)	9.11% (7.13%)	12.15% (0.07%)	<b>22.81%</b> <b>(1.23%)</b>
US	7.35% (0%)	6.38% (0%)	65.08% (0.12%)	32.66% (0.25%)	22.13% (0.36%)	14.94% (4.77%)	11.58% (0.07%)	17.34% (5.61%)	9.81% (25.83%)	10.4% (9.17%)	<b>19.77%</b> <b>(4.62%)</b>
UK	19% (15.16%)	22.96% (6.83%)	69.49% (0.18%)	46.82% (5.78%)	26.64% (7.74%)	18.44% (5.03%)	9.38% (0.01%)	16.23% (0.02%)	14.79% (0.22%)	20.89% (0.08%)	<b>26.46%</b> <b>(4.11%)</b>
Japan	10.12% (0.76%)	3.73% (0.02%)	41.04% (0.28%)	43.49% (1.04%)	22.62% (0.62%)	12.87% (0.11%)	19.49% (0.03%)	11.58% (0.03%)	3.39% (0.01%)	5.76% (0.01%)	<b>17.41%</b> <b>(0.29%)</b>
<b>Emerging</b>	31.64% (0.43%)	38.19% (0.57%)	75.18% (0.63%)	50.2% (1.66%)	23.89% (1.71%)	34.06% (4.92%)	23.63% (0.07%)	18.35% (0.1%)	15.51% (0.04%)	20.47% (0.06%)	<b>33.11%</b> <b>(1.02%)</b>
Russia	44.72% (9.77%)	14.1% (6.37%)	53.41% (1.04%)	104.76% (12.61%)	31.66% (5.06%)	39.9% (3.35%)	32.95% (0.07%)	25.2% (0.1%)	60.47% (0.13%)	42.05% (0.2%)	<b>44.92%</b> <b>(3.87%)</b>
China	54.69% (2.01%)	74.44% (15.23%)	75.41% (0.86%)	46.09% (0.36%)	22.33% (0.19%)	25.56% (3.19%)	19.85% (0.01%)	14.43% (0.03%)	17.94% (0.05%)	49.18% (0.27%)	<b>39.99%</b> <b>(2.22%)</b>
Brazil	57.35% (10.46%)	68.62% (3.21%)	87.29% (2.16%)	57.1% (8.46%)	23.59% (7.2%)	13.57% (3.13%)	16.7% (0.01%)	29.39% (0.08%)	28.77% (0.37%)	43.54% (0.09%)	<b>42.59%</b> <b>(3.52%)</b>
<b>Frontier</b>	21.22% (1.96%)	28.95% (0.56%)	108.66% (3.44%)	33.05% (14.01%)	21.13% (11.5%)	15.07% (2.29%)	16.44% (0.14%)	15.11% (0.16%)	18.46% (0.73%)	20.69% (0.73%)	<b>29.88%</b> <b>(3.55%)</b>
Estonia	30.96% (2.65%)	30.7% (0.04%)	99.4% (2.55%)	63.69% (3.78%)	49.29% (2.38%)	23.41% (0.3%)	20.84% (0.05%)	14.3% (0.12%)	27.51% (2.27%)	8% (0.04%)	<b>36.81%</b> <b>(1.42%)</b>
Morocco	50% (2.44%)	29.76% (20.98%)	52.44% (6.93%)	18.75% (1.65%)	17.55% (0.1%)	6.13% (0.02%)	15.83% (0.08%)	17.1% (0.24%)	7.59% (0.01%)	3.75% (0.01%)	<b>21.89%</b> <b>(3.25%)</b>
Jordan	41.88% (8.19%)	21.65% (0.28%)	65.41% (0.91%)	16.62% (1.38%)	20.99% (0.56%)	17.82% (0.26%)	11.97% (0.06%)	11.41% (0.01%)	7.3% (0.03%)	7.93% (0.03%)	<b>22.3%</b> <b>(1.17%)</b>
<b>Average</b>	<b>31.82%</b> <b>(4.49%)</b>	<b>29.6%</b> <b>(4.51%)</b>	<b>71.45%</b> <b>(1.62%)</b>	<b>45.82%</b> <b>(4.32%)</b>	<b>25.28%</b> <b>(3.17%)</b>	<b>20.33%</b> <b>(2.29%)</b>	<b>18.26%</b> <b>(0.05%)</b>	<b>16.94%</b> <b>(0.81%)</b>	<b>18.39%</b> <b>(3.07%)</b>	<b>20.4%</b> <b>(0.9%)</b>	<b>29.83%</b> <b>(2.52%)</b>

**Note:** The table reports the average OOS annualized returns of twelve cross-validated portfolios for IS of two years and OOS of one month after transaction costs (rolling-forward by one month). In the parentheses, I report the average percentage of cross-validated rules from the total pool of rules. For example, Table 2 reports that 2628 rules ( $0.124 \times 21195$ ) survive on average in the case of frontier markets (2006). This table estimates that from those rules, 52 rules ( $0.0196 \times 2628$ ) are surviving both in the IS and OOS and they generate an average OOS annualized return of 21.22%.

Table 4.12 Annualized Returns Based on the cross-validated surviving rules (IS of 2 Years and OOS 6 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	13.22% (0.02%)	8.9% (0.03%)	45.74% (0.08%)	22.7% (0.73%)	12.31% (0.45%)	12.04% (0.07%)	10.16% (0.02%)	8.82% (3.21%)	5.13% (10.2%)	10.06% (0.07%)	<b>14.91%</b> <b>(1.49%)</b>
US	4.97% (0%)	7.09% (0.01%)	50.45% (0.03%)	20.32% (0.18%)	13.2% (0.25%)	11.41% (0.07%)	9.18% (0.07%)	13.54% (7.34%)	7.65% (27.92%)	9.06% (4.63%)	<b>14.69%</b> <b>(4.05%)</b>
UK	16.89% (17.58%)	15.15% (3.46%)	57.18% (0.19%)	30.06% (4.59%)	17.02% (7.38%)	11.72% (0.21%)	7.73% (0.01%)	12.62% (0.03%)	10.12% (0.1%)	15.6% (0.07%)	<b>19.41%</b> <b>(3.36%)</b>
Japan	1.65% (0.02%)	1.17% (0.01%)	40.34% (0.17%)	26.61% (1.12%)	12.19% (0.48%)	7.81% (0.12%)	10.42% (0.04%)	5.66% (0.03%)	2.52% (0.02%)	5.63% (0.02%)	<b>11.4%</b> <b>(0.2%)</b>
<b>Emerging</b>	19.85% (0.46%)	31.44% (0.77%)	68.91% (0.33%)	28.07% (1.49%)	14.62% (3.44%)	26.61% (1.8%)	17.52% (0.08%)	11.14% (0.12%)	11.06% (0.03%)	16.32% (0.05%)	<b>24.55%</b> <b>(0.86%)</b>
Russia	23.72% (9.25%)	10% (5.9%)	60.63% (1.2%)	59.59% (10.61%)	21.97% (4.92%)	21.17% (3.3%)	18.7% (0.06%)	14.59% (0.08%)	42.1% (0.11%)	27.76% (0.22%)	<b>30.02%</b> <b>(3.56%)</b>
China	36.6% (2.42%)	56.76% (14.35%)	54.33% (0.7%)	24.73% (0.38%)	12.09% (1.74%)	14.03% (0.08%)	8.3% (0.01%)	9.2% (0.02%)	17.33% (0.06%)	30.88% (0.28%)	<b>26.42%</b> <b>(2%)</b>
Brazil	26.05% (6.44%)	50.74% (5.66%)	57.57% (2.61%)	33.43% (10.36%)	13.81% (5.86%)	12.13% (0.06%)	10.35% (0.01%)	20.98% (0.13%)	18.28% (0.3%)	49.05% (0.1%)	<b>29.24%</b> <b>(3.15%)</b>
<b>Frontier</b>	16.76% (1.61%)	18.92% (0.52%)	86.73% (2.72%)	19.73% (13.14%)	12.41% (13.1%)	9.47% (0.73%)	11.17% (0.14%)	13.55% (0.23%)	10.2% (0.46%)	16.49% (0.54%)	<b>21.54%</b> <b>(3.32%)</b>
Estonia	23.28% (1.38%)	25.66% (0.05%)	78.2% (1.96%)	43.24% (4.44%)	23.12% (2.67%)	12.24% (0.25%)	14.04% (0.07%)	6.54% (0.11%)	17.91% (1.56%)	7.16% (0.04%)	<b>25.14%</b> <b>(1.25%)</b>
Morocco	35.66% (2.28%)	22.89% (25.24%)	34.85% (5.99%)	9.48% (1.34%)	12.06% (0.07%)	4.82% (0.02%)	9.27% (0.1%)	11.46% (0.07%)	1.84% (0.01%)	3.39% (0.02%)	<b>14.57%</b> <b>(3.51%)</b>
Jordan	26.1% (7.34%)	14.5% (0.24%)	54% (1.05%)	10.79% (1.03%)	13.75% (0.58%)	10.64% (0.23%)	5.9% (0.04%)	5.08% (0.01%)	6.5% (0.02%)	5.79% (0.03%)	<b>15.31%</b> <b>(1.06%)</b>
<b>Average</b>	<b>20.4%</b> <b>(4.07%)</b>	<b>21.94%</b> <b>(4.69%)</b>	<b>57.41%</b> <b>(1.42%)</b>	<b>27.4%</b> <b>(4.12%)</b>	<b>14.88%</b> <b>(3.41%)</b>	<b>12.84%</b> <b>(0.58%)</b>	<b>11.06%</b> <b>(0.06%)</b>	<b>11.1%</b> <b>(0.95%)</b>	<b>12.55%</b> <b>(3.4%)</b>	<b>16.43%</b> <b>(0.51%)</b>	<b>20.6%</b> <b>(2.32%)</b>

**Note:** The table reports the average OOS annualized returns of twelve cross-validated portfolios for IS of two years and OOS of one month after transaction costs (rolling-forward by one month). In the parentheses, I report the average percentage of cross-validated rules from the total pool of rules. For example, Table 2 reports that 2628 rules ( $0.124 \times 21195$ ) survive on average in the case of frontier markets (2006). This table estimates that from those rules, 42 rules ( $0.0161 \times 2628$ ) are surviving both in the IS and OOS and they generate an average OOS annualized return of 16.76%.



Table 4.13: Financial Stress Performance

Market	Period	Financial Stress	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
US	IS of 2 Years - OOS 1 Month	High	-	-9.96% (-0.59)	22.72% (0.77)	-3.64% (-0.19)	-18.27% (-2.18)	-0.57% (-0.17)	-5.08% (-2.16)	-	-	-3.86% (-0.75)	-2.67% (-0.75)
		Low	1.81% (0.2)	-3.36% (-0.32)	-	-	-0.28% (-0.02)	-2.42% (-0.39)	0.53% (0.18)	7.02% (1.31)	5.73% (0.77)	-5.41% (-0.71)	0.45% (0.13)
Other Developed		High	-	-0.72% (-0.06)	10.63% (0.41)	4.14% (0.25)	-16.24% (-2.88)	-2.15% (-0.55)	-4.86% (-1.81)	-	-	-0.37% (-0.07)	-1.37% (-0.59)
		Low	-3.66% (-0.54)	-6.02% (-0.57)	-	-	-33.23% (-3.59)	0.61% (0.16)	0.11% (0.05)	0.35% (0.13)	-2.07% (-0.42)	-1.53% (-0.85)	-5.68% (-0.7)
Emerging		High	-	-13.42% (-4.45)	43.52% (1.1)	1.78% (0.11)	-20.32% (-3.13)	-4.89% (-0.69)	3% (0.74)	26.12% (4.45)	-4.83% (-4.01)	-1.48% (-0.37)	3.28% (-0.69)
		Low	1.75% (0.15)	-0.82% (-0.06)	-	7.8% (0.7)	-4.37% (-0.66)	-1.87% (-0.36)	-3.65% (-0.94)	-3.94% (-1.45)	-3.61% (-1.21)	0.48% (0.29)	-0.91% (-0.39)
US	IS of 2 Years - OOS 3 Months	High	-	-7.54% (-0.44)	34.43% (1.12)	-14% (-0.89)	-16.52% (-1.78)	-10.15% (-1.41)	-1.75% (-1.81)	-	-	-2.11% (-1.37)	-2.52% (-0.94)
		Low	4.28% (0.48)	-2.28% (-0.26)	-	-	5.43% (0.75)	0.39% (0.06)	-0.72% (-0.27)	5.41% (1.03)	6.43% (0.86)	-5.09% (-0.69)	1.73% (0.25)
Other Developed		High	-	-0.9% (-0.06)	24.85% (0.99)	-5.42% (-0.4)	-13.37% (-2.23)	-2.42% (-0.53)	-3.85% (-1.49)	-	-	2.95% (1.7)	0.26% (-0.29)
		Low	-4.54% (-0.62)	-4.91% (-0.48)	-	-	-	-	-1.58% (-0.76)	-1.35% (-0.46)	0.2% (0.04)	-2.7% (-1.26)	-2.48% (-0.59)
Emerging		High	-	-37.03% (-4.22)	15.33% (0.49)	-8.44% (-0.67)	-11.01% (-1.97)	-0.71% (-0.12)	-3.53% (-0.79)	-	-2.5% (-1.18)	-2.07% (-0.51)	-6.24% (-1.12)
		Low	-4.31% (-0.42)	6.45% (0.45)	-	6.49% (0.62)	-4.64% (-0.7)	0.55% (0.09)	-0.76% (-0.25)	-0.38% (-0.11)	-4.96% (-1.9)	1.1% (0.45)	-0.05% (-0.20)
US	IS of 2 Years - OOS 6 Months	High	-	-11.35% (-0.7)	26.57% (0.9)	-8.53% (-0.62)	-13.35% (-1.36)	-6.68% (-0.81)	-1.1% (-1.27)	-	-	-	-2.41% (-0.64)
		Low	0.53% (0.06)	2.57% (0.23)	-	7.78% (0.78)	6.53% (1.59)	-3.35% (-0.43)	0.98% (0.35)	4.77% (0.95)	5.66% (0.75)	-6.46% (-0.95)	2.11% (0.37)
Other Developed		High	-	-0.42% (-0.03)	13.92% (0.57)	-2.48% (-0.22)	-9.35% (-1.49)	-5.75% (-1.16)	-4.53% (-2.14)	-	-	1.34% (0.54)	-1.04% (-0.56)
		Low	1.05% (0.14)	-3.33% (-0.34)	-	-	-	-	-0.61% (-0.25)	0.74% (0.26)	-0.82% (-0.17)	-1.79% (-0.77)	-0.79% (-0.19)
Emerging		High	-	-15.65% (-1.52)	18.1% (0.57)	2.91% (0.23)	-	-5.57% (-0.88)	-1.23% (-0.24)	-	0.35% (0.15)	-1.87% (-0.45)	-0.42% (-0.3)
		Low	-1.43% (-0.15)	15.56% (1.03)	-	-1.46% (-0.16)	-2.69% (-0.41)	-3.17% (-0.49)	2.99% (1.18)	-0.65% (-0.18)	-3.63% (-1.32)	-	0.69% (-0.06)

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of the portfolios generated in Section 4.5.1. High and low corresponds to high and low financial stress conditions as reported by the OFR stress indexes. – indicates that for this year and market there was no period with high (or low) financial stress.

Table 5.1: List of GARCH, SV, EWMA, and HAR volatility models.

Model	Definition	Equation
ARCH	$\sigma_t^2 = \omega + \sum_{u=1}^q a_u \varepsilon_{t-u}^2$	(5.3)
GARCH/ GARCH-MA	$\sigma_t^2 = \omega + \sum_{u=1}^q a_u \varepsilon_{t-u}^2 + \sum_{r=1}^p \beta_r \sigma_{t-r}^2$	(5.4)
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{u=2}^q a_u (\varepsilon_{t-u}^2 - \varepsilon_{t-1}^2) + \sum_{r=1}^p \beta_r (\sigma_{t-r}^2 - \varepsilon_{t-1}^2)$	(5.5)
Taylor- Schwert	$\sigma_t = \omega + \sum_{u=1}^q a_u  \varepsilon_{t-u}  + \sum_{r=1}^p \beta_r \sigma_{t-r}$	(5.6)
A-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^q [a_u \varepsilon_{t-u}^2 + \eta_u \varepsilon_{t-u}]^2 + \sum_{r=1}^p \beta_r \sigma_{t-r}^2$	(5.7)
NA-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^q a_u (\varepsilon_{t-u} + \eta_u \sigma_{t-u})^2 + \sum_{r=1}^p \beta_r \sigma_{t-r}^2$	(5.8)
TGARCH	$\sigma_t = \omega + \sum_{u=1}^q a_u [(1 - \eta_u) \varepsilon_{t-u}^+ + (1 + \eta_u) \varepsilon_{t-u}^-]^2 + \sum_{r=1}^p \beta_r \sigma_{t-r}$	(5.9)
GJR-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^q [a_u + \eta_u I_{\{\varepsilon_{t-u} < 0\}}] \varepsilon_{t-u}^2 + \sum_{r=1}^p \beta_r \sigma_{t-r}^2$	(5.10)
log-GARCH	$\log(5. \sigma_t) = \omega + \sum_{u=1}^q a_u  e_{t-u}  + \sum_{r=1}^p \beta_r \log(5. \sigma_{t-r})$	(5.11)
EGARCH	$\log(5. \sigma_t^2) = \omega + \sum_{u=1}^q [a_u e_{t-u} + \eta_u (5.  e_{t-u}  - E e_{t-u} )] + \sum_{r=1}^p \beta_r \log(5. \sigma_{t-r}^2)$	(5.12)
NGARCH	$\sigma_t^\delta = \omega + \sum_{u=1}^q a_u  \varepsilon_{t-u} ^\delta + \sum_{r=1}^p \beta_r \sigma_{t-r}^\delta$	(5.13)
APARCH	$\sigma_t^\delta = \omega + \sum_{u=1}^q a_u [ \varepsilon_{t-u}  - \eta_u \varepsilon_{t-u}]^\delta + \sum_{r=1}^p \beta_r \sigma_{t-r}^\delta$	(5.14)
FI-GARCH	$[1 - a(L) - \beta(L)](1 - L)^d \varepsilon_t^2 = \omega + [1 - \beta(L)] \zeta_t$	(5.15)
SV/ SV-MA	$h_t = \mu_h + \sum_{r=1}^p \phi_r [h_{t-r} - \mu_h] + e_t$	(5.16)
SV-L	$h_t = \mu_h + \sum_{u=1}^q [a_u e_{t-u} + \eta_u (5.  e_{t-u}  - E e_{t-u} )] + \sum_{r=1}^p \beta_r [h_{t-r} - \mu_h] + e_t$	(5.17)
EWMA	$\sigma_t^2 = v \sigma_{t-1}^2 + (1 - v) \varepsilon_{t-1}^2$	(5.18)
HAR	$\tilde{\sigma}_t^d = \omega + \beta_d \tilde{\sigma}_{t-1}^d + \beta_w \tilde{\sigma}_{t-1}^w + \beta_{mn} \tilde{\sigma}_{t-1}^{mn} + \varepsilon_t$	(5.19)
log-HAR	$\log(5. \tilde{\sigma}_t^d) = \omega + \beta_d \log(5. \tilde{\sigma}_{t-1}^d) + \beta_w \log(5. \tilde{\sigma}_{t-1}^w) + \beta_{mn} \log(5. \tilde{\sigma}_{t-1}^{mn}) + \varepsilon_t$	(5.20)

Table 5.2: Summary Statistics of Log Returns

	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD
<b>Observations</b>	1877	1877	1877	1509	1516	1871
<b>Mean (%)</b>	-0.004	-0.008	0.02	0.047	0.021	-0.01
<b>Median (%)</b>	-0.009	-0.006	0.024	0.049	0.044	-0.003
<b>Maximum (%)</b>	3.036	2.849	3.496	3.876	3.515	4.666
<b>Minimum (%)</b>	-2.334	-8.429	-3.697	-3.64	-4.780	-9.363
<b>Std. Dev. (%)</b>	0.491	0.497	0.554	0.729	0.863	0.878
<b>Skewness</b>	0.279	-2.473	-0.222	-0.274	-0.161	-0.68
<b>Excess Kurtosis</b>	3.36	45.114	4.336	2.321	2.262	10.42
<b>JB</b>	907.212*	161086.828*	1485.565*	357.574*	329.717*	8607.565*
<b>Q (20)</b>	32.22*	25.782	15.907	26.061	41.636*	19.858
<b>ADF</b>	-30.646*	-42.391*	-43.135*	-39.507*	-19.189*	-44.118*
<b>P-P</b>	-45.057*	-42.391*	-43.135*	-39.507*	-39.023*	-44.118*

**Note:** The JB statistic tests whether the skewness and kurtosis of a sample dataset match the normal distribution. Q (20) is the Ljung–Box statistic which tests if the data is distributed independently. Serial correlation of order up to the 20<sup>th</sup> is considered. ADF and P-P are the statistics of the augmented Dickey-Fuller and Phillips-Perron unit root tests respectively. The lag length for the unit root tests is set based on the lowest Akaike Information Criteria (AIC) value. \* indicates rejection at the 5% significance level.

Table 5.3: Loss Functions

Type	Definition	Equation
<b>MAE<sub>1</sub></b>	$\mathcal{L}_{MAE_1} = K^{-1} \sum_{t=1}^K  \sigma_t - \hat{\sigma}_t $	(5.22)
<b>MAE<sub>2</sub></b>	$\mathcal{L}_{MAE_2} = K^{-1} \sum_{t=1}^K  \sigma_t^2 - \hat{\sigma}_t^2 $	(5.23)
<b>MSE<sub>1</sub></b>	$\mathcal{L}_{MSE_1} = K^{-1} \sum_{t=1}^K (\sigma_t - \hat{\sigma}_t)^2$	(5.24)
<b>MSE<sub>2</sub></b>	$\mathcal{L}_{MSE_2} = K^{-1} \sum_{t=1}^K (\sigma_t^2 - \hat{\sigma}_t^2)^2$	(5.25)
<b>R<sup>2</sup>LOG</b>	$\mathcal{L}_{R^2_{LOG}} = K^{-1} \sum_{t=1}^K [\log(\sigma_t^2 \hat{\sigma}_t^{-2})]^2$	(5.26)
<b>QLIKE</b>	$\mathcal{L}_{QLIKE} = K^{-1} \sum_{t=1}^K (\log(\hat{\sigma}_t^2) + \sigma_t^2 \hat{\sigma}_t^{-2})$	(5.27)

**Note:** The conditional volatilities estimated by the forecasting models are presented by  $\hat{\sigma}_t$  and compared to the actual ones  $\sigma_t$ .  $K$  is the number of forecasting points, set by the number of trading days in each calendar year. I choose the  $K$  adaptively based on the dataset rather than a fixed quantity.

Table 5.4: Variation in Number of True Discoveries

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	4	469	969	481	716	527.8
	GARCH (1,1)	12	452	2	496	96	211.6
	PRC 90	1	26	1	60	96	36.8
GBP/USD	ARCH (1)	265	1	726	175	114	256.2
	GARCH (1,1)	96	3	4	110	126	67.8
	PRC 90	96	1	1	97	113	61.6
USD/JPY	ARCH (1)	399	1	1	944	332	335.4
	GARCH (1,1)	1	440	1	534	274	250
	PRC 90	7	1	1	2	10	4.2
DJIA	ARCH (1)	1	34	105	32	33	41
	GARCH (1,1)	1	118	304	33	38	98.8
	PRC 90	91	103	105	42	40	76.2
FTSE 100	ARCH (1)	190	2	1	1	4	39.6
	GARCH (1,1)	203	48	360	1	19	126.2
	PRC 90	91	37	106	50	64	69.6
XAU/USD	ARCH (1)	326	3	471	589	2	278.2
	GARCH (1,1)	414	3	448	24	211	220
	PRC 90	56	1	1	1	2	12.2

**Note:** The table presents the size of the rejection set for different markets based on the  $MSE_1$  benchmark. The  $DFDR^+$  always rejects the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table 5.5: Innovation Distribution Survival Rate Across the Markets.

Asset	Benchmark	Gaussian	t	Skewed t	GED	No dist.
EUR/USD	ARCH (1)	36.56%	34.95%	36.13%	<b>38.58%</b>	3.33%
	GARCH (1,1)	14.78%	14.41%	<b>14.78%</b>	14.44%	1.67%
	PRC 90	2.63%	2.31%	<b>3.01%</b>	2.22%	0.00%
GBP/USD	ARCH (1)	17.69%	17.96%	16.34%	<b>19.01%</b>	1.67%
	GARCH (1,1)	4.52%	4.73%	<b>5.16%</b>	4.38%	0.00%
	PRC 90	4.19%	<b>4.35%</b>	<b>4.35%</b>	4.20%	0.00%
USD/JPY	ARCH (1)	24.68%	22.80%	17.42%	<b>27.78%</b>	5.56%
	GARCH (1,1)	17.85%	17.63%	14.78%	<b>19.14%</b>	1.39%
	PRC 90	0.48%	0.11%	0.05%	<b>0.56%</b>	0.00%
DJIA	ARCH (1)	2.96%	<b>3.55%</b>	2.85%	1.91%	0.00%
	GARCH (1,1)	7.26%	<b>7.53%</b>	6.29%	5.99%	1.39%
	PRC 90	5.38%	<b>6.34%</b>	5.65%	3.58%	0.00%
FTSE 100	ARCH (1)	2.69%	<b>2.96%</b>	2.69%	2.59%	0.28%
	GARCH (1,1)	8.49%	8.98%	<b>9.68%</b>	7.16%	2.78%
	PRC 90	4.95%	5.27%	<b>5.27%</b>	3.64%	0.28%
XAU/USD	ARCH (1)	17.10%	18.66%	<b>24.46%</b>	16.42%	1.39%
	GARCH (1,1)	12.37%	15.11%	<b>20.91%</b>	12.22%	0.56%
	PRC 90	0.54%	0.70%	<b>1.67%</b>	0.43%	0.00%
Average		10.28%	10.46%	<b>10.64%</b>	10.24%	1.13%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 36.56% means that on average 136 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is MSE<sub>1</sub>. The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.

Table 5.6: Classes Survival Rates

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	10.83%	1.67%	14.86%	17.50%	24.44%	18.75%	14.68%
GARCH	19.17%	2.99%	14.79%	3.13%	8.61%	13.26%	10.32%
IGARCH	<b>43.19%</b>	<b>65.14%</b>	<b>28.82%</b>	7.01%	1.94%	<b>23.47%</b>	<b>28.26%</b>
Taylor/Schwert	19.65%	10.83%	17.43%	11.81%	6.53%	19.86%	14.35%
A-GARCH	18.96%	2.92%	14.10%	0.35%	0.42%	7.01%	7.29%
NA-GARCH	19.44%	2.99%	13.19%	0.35%	0.56%	6.74%	7.21%
TGARCH	26.46%	10.07%	15.63%	0.35%	0.42%	8.19%	10.19%
GJR-GARCH	17.22%	2.78%	11.88%	0.00%	0.00%	6.46%	6.39%
log-GARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NGARCH	21.46%	11.67%	18.06%	2.85%	1.81%	17.01%	12.14%
APARCH	21.39%	10.56%	18.54%	0.21%	0.56%	11.67%	10.49%
FI-GARCH	12.15%	2.08%	15.21%	13.82%	17.08%	22.43%	13.80%
GARCH-MA	18.96%	3.13%	16.25%	3.75%	10.97%	17.50%	11.76%
SV	11.11%	1.67%	15.56%	<b>29.44%</b>	<b>26.67%</b>	18.89%	17.22%
SV-MA	11.11%	2.22%	15.56%	28.33%	26.11%	17.78%	16.85%
SV-L	21.94%	8.06%	0.00%	0.83%	0.00%	0.00%	5.14%
RM	2.00%	0.67%	2.78%	0.56%	1.33%	0.78%	1.35%
HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
log-HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 10.83% means that on average 5 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is  $MSE_1$  and the value in bold shows the maximum of each column.

Table A.1: EUR/USD Trading Performance – Sharpe Ratio

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	-0.22	-0.08	-0.02	3	SA (5)	0.19	0.34	-0.22
	SA (10)	0.20	-0.18	0.06		SA (10)	-0.22	-0.07	-0.37
	SA (15)	-0.14	-0.11	-0.21		SA (15)	-0.12	-0.28	-0.20
	NB (5)	0.18	0.14	0.18		NB (5)	0.42	0.40	0.44
	NB (10)	0.33	0.29	0.35		NB (10)	0.25	0.36	0.47
	NB (15)	0.30	0.26	0.31		NB (15)	0.27	0.37	0.49
	DMA (5)	0.44	0.46	0.48		DMA (5)	<b>0.80</b>	0.61	0.54
	DMA (10)	<b>0.62</b>	0.44	0.58		DMA (10)	0.71	0.58	0.72
	DMA (15)	0.57	0.59	0.55		DMA (15)	0.62	<b>0.69</b>	0.65
	DMS (5)	0.50	<b>0.67</b>	0.52		DMS (5)	0.57	0.49	0.61
	DMS (10)	0.48	0.55	<b>0.58</b>		DMS (10)	0.69	0.54	0.63
	DMS (15)	0.47	0.49	0.26		DMS (15)	0.68	0.56	0.69
	BNN (5)	0.59	0.57	0.48		BNN (5)	0.70	0.58	0.71
	BNN (10)	0.60	0.63	0.52		BNN (10)	0.77	0.66	<b>0.74</b>
	BNN (15)	0.59	0.58	0.55		BNN (15)	0.64	0.68	0.72
2	RVM	0.50	0.50	0.50	4	RVM	0.63	0.63	0.63
	SA (5)	0.09	-0.07	0.12		SA (5)	0.17	-0.08	-0.22
	SA (10)	0.19	-0.11	0.14		SA (10)	-0.20	0.20	-0.29
	SA (15)	0.08	0.22	0.17		SA (15)	-0.12	0.18	-0.10
	NB (5)	0.18	0.28	0.13		NB (5)	0.36	-0.24	0.25
	NB (10)	0.26	0.26	0.18		NB (10)	0.38	0.26	0.11
	NB (15)	0.17	0.26	0.15		NB (15)	0.41	0.33	0.28
	DMA (5)	0.32	<b>0.41</b>	0.53		DMA (5)	0.64	0.47	0.39
	DMA (10)	0.44	0.30	0.55		DMA (10)	0.56	<b>0.52</b>	0.42
	DMA (15)	0.48	0.32	0.59		DMA (15)	0.65	0.49	0.37
	DMS (5)	0.35	0.29	0.53		DMS (5)	0.40	0.38	0.38
	DMS (10)	0.37	0.26	0.55		DMS (10)	0.73	0.41	0.40
	DMS (15)	0.44	0.24	0.49		DMS (15)	0.74	0.44	0.42
	BNN (5)	0.36	0.22	0.52		BNN (5)	0.69	0.41	0.54
	BNN (10)	<b>0.51</b>	0.34	0.48		BNN (10)	0.65	0.50	<b>0.70</b>
	BNN (15)	0.40	0.33	<b>0.62</b>		BNN (15)	<b>0.76</b>	0.51	0.59
	RVM	0.31	0.31	0.31		RVM	0.47	0.47	0.47

**Note:** The table presents the annualized Sharpe Ratio for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.



Table A.2: GBP/USD Trading Performance – Sharpe Ratio

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	0.30	-0.40	-0.03	3	SA (5)	-0.07	0.14	-0.21
	SA (10)	0.12	-0.08	-0.39		SA (10)	0.23	-0.24	0.27
	SA (15)	-0.15	-0.02	-0.26		SA (15)	0.19	-0.13	0.31
	NB (5)	0.32	0.21	0.13		NB (5)	0.27	0.22	0.26
	NB (10)	0.33	0.28	0.34		NB (10)	0.31	0.39	0.42
	NB (15)	0.38	0.39	0.20		NB (15)	0.37	0.43	0.17
	DMA (5)	0.43	0.51	0.41		DMA (5)	0.64	0.45	0.46
	DMA (10)	0.38	0.57	<b>0.65</b>		DMA (10)	0.62	0.49	0.51
	DMA (15)	0.65	0.63	0.36		DMA (15)	<b>0.79</b>	<b>0.81</b>	0.48
	DMS (5)	0.44	0.40	0.42		DMS (5)	0.64	0.32	0.39
	DMS (10)	0.37	0.59	0.38		DMS (10)	0.62	0.46	0.57
	DMS (15)	0.56	0.48	0.32		DMS (15)	0.68	0.72	0.57
	BNN (5)	0.42	0.51	0.63		BNN (5)	0.53	0.60	0.54
	BNN (10)	0.61	0.50	0.45		BNN (10)	0.48	0.63	0.32
	BNN (15)	<b>0.75</b>	<b>0.66</b>	0.38		BNN (15)	0.71	0.70	<b>0.68</b>
2	RVM	0.41	0.41	0.41	4	RVM	0.39	0.39	0.39
	SA (5)	0.09	0.11	-0.29		SA (5)	-0.05	0.18	0.05
	SA (10)	0.20	0.14	0.14		SA (10)	0.18	0.14	0.27
	SA (15)	0.24	0.04	-0.11		SA (15)	0.09	0.07	-0.04
	NB (5)	0.26	0.33	0.23		NB (5)	0.26	0.26	0.25
	NB (10)	0.28	0.41	0.20		NB (10)	0.29	0.37	0.34
	NB (15)	0.19	0.43	0.15		NB (15)	0.31	0.48	0.37
	DMA (5)	0.37	0.39	0.56		DMA (5)	0.48	0.51	0.59
	DMA (10)	0.50	0.42	<b>0.61</b>		DMA (10)	0.51	0.62	<b>0.64</b>
	DMA (15)	0.72	0.63	0.49		DMA (15)	<b>0.83</b>	0.78	0.52
	DMS (5)	0.45	0.31	0.57		DMS (5)	0.36	0.41	0.43
	DMS (10)	0.57	0.33	0.48		DMS (10)	0.42	0.50	0.49
	DMS (15)	0.69	0.58	0.34		DMS (15)	0.67	0.66	0.52
	BNN (5)	0.61	0.45	0.53		BNN (5)	0.52	0.64	0.57
	BNN (10)	0.65	<b>0.67</b>	0.57		BNN (10)	0.57	0.61	0.56
	BNN (15)	<b>0.76</b>	0.61	0.49		BNN (15)	0.64	<b>0.80</b>	0.36
	RVM	0.43	0.43	0.43		RVM	0.48	0.48	0.48

Note: The table presents the annualized Sharpe Ratio for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.

Table A.3: USD/JPY Trading Performance – Sharpe Ratio

Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio	Exercise	Measure	Accuracy	Profit-ability	Sharpe Ratio
1	SA (5)	-0.30	0.31	0.32	3	SA (5)	-0.21	-0.19	-0.31
	SA (10)	-0.09	0.4	0.45		SA (10)	-0.15	-0.16	0.10
	SA (15)	-0.21	0.15	0.40		SA (15)	0.03	-0.12	-0.17
	NB (5)	0.26	0.42	0.42		NB (5)	0.17	0.25	-0.19
	NB (10)	0.22	0.29	0.46		NB (10)	0.39	0.07	0.04
	NB (15)	0.34	0.25	0.34		NB (15)	0.27	0.26	0.06
	DMA (5)	0.38	0.64	<b>0.76</b>		DMA (5)	0.45	0.48	<b>0.63</b>
	DMA (10)	0.47	0.48	0.61		DMA (10)	0.46	0.45	0.59
	DMA (15)	0.53	<b>0.74</b>	0.65		DMA (15)	0.58	0.67	0.41
	DMS (5)	0.60	0.49	0.69		DMS (5)	0.46	0.40	0.28
	DMS (10)	0.55	0.40	0.52		DMS (10)	0.42	0.33	0.47
	DMS (15)	0.43	0.63	0.46		DMS (15)	0.53	0.68	0.30
	BNN (5)	0.46	0.55	0.50		BNN (5)	<b>0.62</b>	0.56	0.45
	BNN (10)	0.54	0.51	0.57		BNN (10)	0.51	0.53	0.62
	BNN (15)	<b>0.68</b>	0.63	0.60		BNN (15)	0.43	<b>0.69</b>	0.50
2	RVM	0.54	0.54	0.54	4	RVM	0.41	0.41	0.41
	SA (5)	-0.25	0.17	0.25		SA (5)	-0.40	-0.24	-0.30
	SA (10)	-0.42	0.25	0.18		SA (10)	-0.23	-0.28	-0.21
	SA (15)	0.06	0.10	0.05		SA (15)	0.09	-0.11	-0.35
	NB (5)	0.24	0.26	0.11		NB (5)	0.18	0.24	-0.16
	NB (10)	0.28	0.14	0.24		NB (10)	0.20	0.20	0.21
	NB (15)	0.36	0.20	0.36		NB (15)	0.23	0.21	0.15
	DMA (5)	0.53	0.56	<b>0.64</b>		DMA (5)	0.76	0.74	0.43
	DMA (10)	0.64	<b>0.73</b>	0.60		DMA (10)	0.55	<b>0.86</b>	0.58
	DMA (15)	<b>0.67</b>	0.58	0.35		DMA (15)	0.63	0.75	0.67
	DMS (5)	0.32	0.47	0.25		DMS (5)	0.59	0.62	0.45
	DMS (10)	0.41	0.33	0.46		DMS (10)	0.49	0.70	0.49
	DMS (15)	0.56	0.62	0.53		DMS (15)	0.60	0.62	0.54
	BNN (5)	0.50	0.70	0.40		BNN (5)	<b>0.71</b>	0.54	0.43
	BNN (10)	0.63	0.53	0.57		BNN (10)	0.62	0.59	<b>0.70</b>
	BNN (15)	0.62	0.64	0.61		BNN (15)	0.69	0.52	0.48
	RVM	0.49	0.49	0.49		RVM	0.56	0.56	0.56

Note: The table presents the annualized Sharpe Ratio for the top rules out of the data-snooping procedure survivors. Three fixed levels (5, 10, and 15) are studied SA, NB, DMA, DMS, DMA and BNN while the RVM is selecting the most relevant rules endogenously. The best rules are selected based on three measures of IS accuracy, profitability, and Sharpe ratio. All returns are after transaction costs. The values in bold correspond to the best performing combination for each criterion and exercise.

**Table B.1: Relevance Vectors**

Betbrain average over 2.5 goals odds (BbAv>2.5)	Betbrain size of handicap (home team) (BBAHh)	Betbrain average Asian handicap home team odds (BbAvAHH)	Betbrain average Asian handicap away team odds (BbAvAHH)	Points of H in the last 3 games when H plays at home (PtH3H)
Points of A in the last 3 games when A plays away (PtA3A)	Number of shots on target of H team in the last 1 game (StH1)	Number of shots on target of A team in the last 1 games (StA1)	Number of corner kicks on target of H team in the last 3 game (CkH3)	Number of corner kicks on target of H team in the last 2 game (CkH2)
Number of corner kicks on target of H team in the last 1 game plays home (CkH1H)	Number of corner kicks on target of A team in the last 2 game plays away (CkA2A)			

**Note:** Team H is the home team and team A is the away team. The parentheses represent the relevant abbreviation of the RV. In total, 12 RVs are selected.

Table B.2: Cluster Characteristics for the Generated Rules

Input variable \ Rule	Rule												
	1	2	3	4	5	6	7	8	9	10	11	12	13
<b>BbAv&gt;2.5</b>	2.136 (0.168)	2.117 (0.128)	1.96 (0.122)	2.034 (0.153)	1.998 (0.173)	2.003 (0.170)	2.093 (0.154)	1.754 (0.160)	2.015 (0.149)	1.767 (0.156)	2.055 (0.135)	2.032 (0.203)	2.134 (0.100)
<b>BbAHh</b>	0.000 (0.797)	-0.502 (0.793)	0.002 (0.793)	0.000 (0.793)	-0.001 (0.795)	0.001 (0.795)	-0.001 (0.797)	-1.502 (0.793)	-0.002 (0.795)	-1.500 (0.794)	-0.500 (0.796)	-0.751 (0.795)	-0.245 (0.795)
<b>BbAvAHH</b>	1.83 (1.419)	1.971 (1.42)	1.82 (1.417)	1.941 (1.421)	1.519 (1.419)	2.05 (1.42)	1.579 (1.419)	1.79 (1.419)	1.7 (1.416)	2.031 (1.427)	1.809 (1.419)	1.84 (1.416)	1.88 (1.42)
<b>BbAvAHA</b>	1.99 (1.715)	1.889 (1.713)	1.9 (1.717)	1.81 (1.715)	2.36 (1.713)	1.741 (1.715)	2.29 (1.715)	2.08 (1.716)	2.099 (1.713)	1.83 (1.714)	2.08 (1.715)	2.04 (1.715)	1.981 (1.716)
<b>PtH3H</b>	4 (1.589)	6.001 (1.593)	3 (1.592)	4.001 (1.588)	0.999 (1.591)	6.999 (1.596)	3 (1.589)	6.001 (1.593)	6.999 (1.589)	6.998 (1.593)	8.999 (1.592)	3.999 (1.594)	4 (1.59)
<b>PtA3A</b>	2.997 (1.587)	1 (1.589)	1.999 (1.591)	6 (1.59)	1.001 (1.591)	5.002 (1.594)	5 (1.592)	3 (1.592)	2.997 (1.587)	0 (1.591)	0.999 (1.591)	0.001 (1.592)	3.003 (1.59)
<b>StH1</b>	4.001 (3.889)	2.999 (3.889)	6 (3.889)	4 (3.89)	6.001 (3.889)	6 (3.89)	4.999 (3.888)	8 (3.889)	7 (3.888)	9 (3.889)	3.999 (3.888)	4 (3.889)	4.001 (3.891)
<b>StA1</b>	5.999 (3.712)	6 (3.712)	3 (3.713)	6.002 (3.713)	7.999 (3.711)	9 (3.714)	6 (3.711)	10 (3.712)	3 (3.711)	5 (3.712)	7 (3.712)	4 (3.713)	5 (3.713)
<b>CkH3</b>	13 (6.187)	15 (6.187)	19 (6.187)	13 (6.187)	13 (6.187)	17 (6.187)	19 (6.187)	16 (6.187)	16.999 (6.187)	19 (6.188)	9 (6.187)	9 (6.187)	21 (6.187)
<b>CkH2</b>	7.001 (4.773)	11 (4.774)	14 (4.773)	10 (4.774)	8.001 (4.773)	9.001 (4.773)	13 (4.773)	12 (4.773)	11 (4.773)	10 (4.774)	5 (4.773)	5.999 (4.773)	16 (4.773)
<b>CkH1H</b>	3.999 (3.358)	6.001 (3.36)	6.001 (3.36)	6 (3.359)	5 (3.358)	4.999 (3.358)	10.001 (3.357)	8 (3.359)	6 (3.358)	6 (3.358)	2.999 (3.358)	4.001 (3.36)	12 (3.358)
<b>CkA2A</b>	7.999 (4.064)	11 (4.067)	10 (4.065)	7.001 (4.068)	9 (4.067)	10 (4.066)	12 (4.065)	8 (4.067)	8.001 (4.067)	8 (4.066)	9.999 (4.065)	7.999 (4.064)	4 (4.065)

**Note:** The values in the Table represent the centre (standard deviation) of the relevant cluster. For instance, consider the first input (BbAv>2.5); to determine which rule an observation belongs to, the membership grade is calculated by the membership function of  $\exp(-(x_{*,1} - 2.136)^2 / (2 \times 0.168^2))$  where  $x_{*,1}$  is the given odd for the Betbrain average for greater than 2.5 goals for the match. The firing strength (weight) of each rule is the product of the membership grades for all inputs.

Table B.3: Regression coefficients for the generated rules

Rule Coefficient	1	2	3	4	5	6	7	8	9	10	11	12	13
<b>Intercept</b>	<b>-2.706</b>	-9.438	10.118	2.951	-10.841	0.159	-4.962	2.463	10.566	-8.081	-12.303	15.935	3.463
<b>BbAv&gt;2.5</b>	<b>1.042</b>	1.68	-2.443	-0.18	0.855	-0.305	0.675	-0.991	0.256	1.118	-0.82	-0.446	4.488
<b>BbAHh</b>	<b>-0.755</b>	-0.196	-0.503	-0.256	-1.064	-0.742	-0.338	-0.431	-0.723	-0.7	-0.402	-1.759	-3.1
<b>BbAvAHH</b>	<b>1.072</b>	2.361	-2.385	-0.962	2.639	-0.251	0.459	-0.191	-4.478	1.548	4.102	-6.554	-3.701
<b>BbAvAHA</b>	<b>0.484</b>	1.055	-1.057	-1.099	1.544	0.119	1.22	-0.353	-0.646	0.952	1.537	-2.011	-1.609
<b>PtH3H</b>	<b>-0.041</b>	-0.141	0.163	0.074	0.043	-0.094	-0.016	-0.096	-0.194	0.074	0.417	0.037	-0.343
<b>PtA3A</b>	<b>0.207</b>	-0.443	0.108	0.087	0.142	0.054	0.225	0.158	0.022	0.171	0.365	-0.743	0.097
<b>StH1</b>	<b>-0.138</b>	0.083	0.049	-0.008	0.022	-0.017	0	0.013	-0.002	0.008	-0.138	0.057	-0.118
<b>StA1</b>	<b>-0.088</b>	0.124	-0.161	0.011	-0.042	0.031	-0.032	0.078	0.156	0.062	-0.083	0.206	-0.105
<b>CkH3</b>	<b>-0.197</b>	0.045	0.139	0.02	0.078	0.033	-0.023	-0.069	0.001	-0.059	-0.041	0.028	-0.079
<b>CkH2</b>	<b>-0.016</b>	-0.028	-0.124	0.043	0.029	-0.058	0.035	0.111	0.038	-0.021	0.016	-0.055	0.03
<b>CkH1H</b>	<b>0.152</b>	-0.005	-0.056	-0.083	0.046	0.065	0	-0.04	-0.147	0.161	0.151	-0.115	0.059
<b>CkA2A</b>	<b>0.074</b>	-0.093	0.05	-0.029	-0.017	0.044	-0.053	-0.025	0.03	0.031	0.061	0.08	-0.005

Table B.4: IS Accuracy

Model	Championship	2006- 2009	2006- 2010	2007- 2011	2008- 2012	2009- 2013	2010- 2014	Average
<b>Game Result</b>	Premiership	87.52%	86.95%	85.94%	83.68%	87.90%	86.54%	86.42%
	La -Liga	87.90%	80.14%	83.27%	84.49%	88.23%	84.38%	84.74%
	Seria A	80.82%	84.06%	79.74%	82.61%	88.12%	83.47%	83.14%
<b>Asian Handicap</b>	Premiership	88.01%	85.22%	83.14%	79.29%	83.36%	84.63%	83.94%
	La -Liga	87.70%	88.34%	84.44%	81.42%	81.11%	79.94%	83.83%
	Seria A	84.47%	79.89%	82.61%	83.05%	81.06%	81.77%	82.14%
<b>Number of Goals</b>	Premiership	86.19%	81.50%	84.55%	83.07%	80.39%	84.48%	83.36%
	La -Liga	85.43%	81.22%	79.39%	81.00%	79.71%	84.19%	81.82%
	Seria A	83.20%	78.32%	84.53%	82.16%	86.92%	82.94%	83.01%

**Note:** All values in the Table represent accuracy ratios

Table B.5: CF Forecasts

Model	Championship	OOS	IS					
			2006-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014
<b>Game Result</b>	Premiership	1	26	25	22	23	21	30
		2	21	18	18	17	23	16
	La -Liga	1	15	60	26	28	43	58
		2	21	23	19	23	47	48
	Seria A	1	25	23	21	25	40	24
		2	49	38	22	19	27	28
<b>Asian Handicap</b>	Premiership	1	29	21	24	28	29	22
		2	26	36	32	24	18	35
	La -Liga	1	27	59	48	39	27	24
		2	48	37	59	45	24	34
	Seria A	1	32	31	24	21	17	31
		2	37	58	21	32	53	35
<b>Number of Goals</b>	Premiership	1	24	19	21	26	24	13
		2	16	36	34	42	76	52
	La -Liga	1	24	51	29	25	59	80
		2	23	48	55	40	22	94
	Seria A	1	27	26	24	28	24	37
		2	30	46	54	29	26	37

**Note:** All values in the Table represent number of games that CF generated forecasts for the respective championship and season. Row 1 corresponds to the first year of the OOS and row 2 to the second year of the OOS. For example, for the first cell on the left corner, 26 is the number of forecasts generated by CF for the 2009-2010 Premiership season and the 21 is the number of forecasts generated by the exact CF model (same specification and rules) for the 2010-2011 season of the same championship. All other models under study are unconditional and generate forecasts for every single game.

Table C.1: Annualized Mean Excess Returns for Quartiles of Different Combination of Sharpe Ratio Levels.

Outperforming SR	Quartile	Underperforming SR					
		-2		-3		-4	
		Outperforming	Underperforming	Outperforming	Underperforming	Outperforming	Underperforming
2	1st	6.50	-8.93	6.80	-16.28	6.60	-24.69
	2nd	16.28	-16.08	16.92	-23.84	16.51	-32.74
	3rd	19.14	-22.55	19.83	-30.71	19.25	-39.62
3	1st	11.51	-8.69	11.37	-16.83	11.43	-24.64
	2nd	25.14	-15.86	24.63	-24.52	24.81	-32.82
	3rd	27.94	-22.53	27.38	-31.05	27.54	-39.55
4	1st	16.44	-8.88	16.32	-16.66	16.64	-24.19
	2nd	33.56	-16.00	33.43	-24.34	34.22	-32.23
	3rd	36.19	-22.47	36.08	-30.97	36.90	-39.13

**Note:** The table reports the quartiles of the distribution of the annualized mean excess return (in percentages) induced by positive and negative Sharpe ratio pairs applied in the Monte Carlo simulations for the out- and under-performing strategies. The pairs are created with the annualized Sharpe ratio for out- and under-performing rules set to 2, 3, 4 and -2, -3, -4 respectively. The quantities presented correspond to the average values over 1000 Monte Carlo simulations. The proportion of rules that are neutrally performing ( $\pi_0$ ), outperforming ( $\pi_A^+$ ) and underperforming ( $\pi_A^-$ ) and are set to 50%, 20% and 30% respectively.



Table C.2 Estimation of Neutral, Positive, and Negative Proportions by the DFDR<sup>+/−</sup> Procedure Versus the Actual Ones.

Outperforming SR	Proportion	Underperforming SR		
		-2	-3	-4
2	$\pi_0 = 50\%$	73.08	62.93	60.97
	$\pi_A^+ = 20\%$	9.39	11.91	11.17
	$\pi_A^- = 30\%$	17.53	25.17	27.86
3	$\pi_0 = 50\%$	66.43	57.09	53.44
	$\pi_A^+ = 20\%$	14.35	15.23	16.09
	$\pi_A^- = 30\%$	19.22	27.68	30.47
4	$\pi_0 = 50\%$	64.35	53.55	50.21
	$\pi_A^+ = 20\%$	16.28	18.08	18.89
	$\pi_A^- = 30\%$	19.37	28.36	30.91

**Note:** The quantities presented correspond to the average values estimated over 1000 Monte Carlo simulations. The proportion of rules that are neutrally performing ( $\pi_0$ ), outperforming ( $\pi_A^+$ ) and underperforming ( $\pi_A^-$ ) and are set to 50%, 20% and 30% respectively. The table provides the estimates when annualized Sharpe ratio for out- and under-performing rules is set to 2, 3, 4 and -2, -3, -4 respectively.

Table C.3: True FDR, Accuracy and the Positive-performing Portfolio Size through Different Methods.

Outperforming SR	Portfolio Type	Underperforming SR								
		-2			-3			-4		
		FDR+	Power	Portfolio size	FDR+	Power	Portfolio size	FDR+	Power	Portfolio size
2	10%-DFDR +	12.92	39.46	1995.71	12.67	42.01	2118.86	12.05	40.31	2014.43
	20%-DFDR +	13.79	39.92	2110.78	15.03	43.3	2369.41	14.65	41.66	2281.92
	5%-RW	0.86	0.01	0.53	0.90	0.01	0.52	0.71	0.01	0.48
	20%-RW	8.42	0.05	3.03	8.28	0.06	3.33	7.11	0.05	2.90
3	10%-DFDR +	8.44	64.74	3076.41	8.00	64.74	3048.45	8.78	62.89	3039.27
	20%-DFDR +	9.54	65.29	3212.88	10.59	66.30	3321.58	11.55	64.62	3343.04
	5%-RW	0.31	0.01	0.52	0.04	0.01	0.60	0.08	0.01	0.57
	20%-RW	3.21	0.07	3.47	2.30	0.07	3.41	2.97	0.07	3.49
4	10%-DFDR +	6.45	82.39	3790.07	6.62	83.57	3879.56	7.83	83.01	3945.84
	20%-DFDR +	7.80	83.35	3948.29	9.53	85.49	4194.29	10.5	84.89	4244.14
	5%-RW	0.00	0.02	0.65	0.01	0.02	0.73	0.00	0.02	0.68
	20%-RW	0.53	0.09	3.87	0.31	0.10	4.38	0.34	0.09	3.73

**Note:** The table reports the FDR+ and accuracy in percentages and the portfolio size (out of 21195). Accuracy is estimated by the ratio of actual outperformers discovered by the underlying procedure. I consider confidence target levels of 10% and 20% for the DFDR+ and benchmark it against the procedure in Romano and Wolf (2005) for the target levels of 5% and 20%. The quantities refer to average values over 1000 Monte Carlo simulations for different combinations pairs when annualized Sharpe ratio for out- and under-performing rules are set to 2, 3, 4 and -2, -3, -4 respectively.

Table C.4: Percentage and Standard Deviation of the DFDR<sup>+-</sup> Procedure Survivors (IS 1 Year).

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	1.37 (2.82)	0.20 (0.20)	1.34 (3.22)	5.44 (4.12)	4.32 (7.72)	2.10 (5.26)	0.42 (0.20)	8.93 (11.28)	7.72 (12.61)	0.45 (0.12)	<b>3.23</b> <b>(4.76)</b>
US	0.02 (0.02)	0.11 (0.12)	0.57 (1.27)	2.53 (2.68)	6.02 (9.25)	2.15 (5.61)	2.06 (5.34)	16.75 (15.89)	16.24 (16.68)	0.32 (0.13)	<b>4.68</b> <b>(5.70)</b>
UK	11.52 (6.91)	14.52 (11.99)	2.75 (6.43)	39.02 (9.20)	13.40 (14.60)	0.57 (0.23)	0.24 (0.07)	2.29 (6.25)	3.17 (8.59)	0.37 (0.11)	<b>8.78</b> <b>(6.44)</b>
Japan	13.19 (13.86)	1.26 (2.59)	2.70 (4.57)	7.01 (2.60)	4.90 (7.97)	0.66 (0.22)	0.49 (0.20)	2.27 (5.30)	0.23 (0.10)	0.44 (0.15)	<b>3.32</b> <b>(3.76)</b>
<b>Emerging</b>	4.87 (9.35)	1.94 (2.38)	1.01 (1.33)	6.74 (5.01)	10.21 (11.65)	0.48 (0.30)	0.61 (0.19)	0.37 (0.21)	0.32 (0.17)	0.53 (0.67)	<b>2.71</b> <b>(3.12)</b>
Russia	31.04 (19.00)	1.92 (3.65)	2.81 (6.59)	33.80 (12.19)	11.24 (12.74)	2.65 (5.77)	0.75 (0.45)	0.94 (0.63)	0.96 (0.38)	1.39 (0.77)	<b>8.75</b> <b>(6.22)</b>
China	7.34 (8.51)	27.46 (23.36)	0.96 (0.61)	2.90 (5.44)	5.68 (9.01)	0.46 (0.70)	0.37 (0.08)	0.91 (0.47)	0.31 (0.25)	5.60 (9.06)	<b>5.20</b> <b>(5.75)</b>
Brazil	33.66 (30.99)	8.85 (12.55)	16.00 (15.17)	34.05 (9.11)	10.67 (11.65)	0.26 (0.20)	0.40 (0.40)	0.97 (0.53)	1.44 (0.61)	1.24 (1.46)	<b>10.75</b> <b>(8.27)</b>
<b>Frontier</b>	9.18 (14.08)	6.12 (8.23)	3.58 (9.53)	38.73 (14.25)	7.03 (8.80)	1.64 (0.80)	0.51 (0.30)	3.05 (5.17)	8.17 (9.53)	2.80 (1.58)	<b>8.08</b> <b>(7.23)</b>
Estonia	0.66 (1.28)	0.83 (0.50)	6.04 (9.17)	7.08 (5.44)	10.04 (10.78)	1.09 (0.98)	0.47 (0.30)	25.02 (8.96)	1.76 (2.35)	1.50 (2.86)	<b>5.45</b> <b>(4.26)</b>
Morocco	19.28 (22.47)	4.67 (4.39)	12.35 (10.54)	2.02 (1.55)	0.31 (0.23)	0.18 (0.06)	1.84 (2.17)	0.22 (0.06)	0.42 (0.16)	0.59 (0.82)	<b>4.19</b> <b>(4.24)</b>
Jordan	0.89 (1.19)	1.25 (1.03)	2.38 (2.98)	4.38 (2.67)	0.39 (0.11)	0.82 (0.96)	0.35 (0.13)	0.26 (0.21)	0.27 (0.10)	0.38 (0.21)	<b>1.14</b> <b>(0.96)</b>
<b>Average</b>	<b>11.08</b> <b>(10.87)</b>	<b>5.76</b> <b>(5.92)</b>	<b>4.37</b> <b>(5.95)</b>	<b>15.31</b> <b>(6.19)</b>	<b>7.02</b> <b>(8.71)</b>	<b>1.09</b> <b>(1.76)</b>	<b>0.71</b> <b>(0.82)</b>	<b>5.17</b> <b>(4.58)</b>	<b>3.42</b> <b>(4.29)</b>	<b>1.30</b> <b>(1.49)</b>	<b>5.52</b> <b>(5.06)</b>

**Note:** The table reports the percentage and standard deviations of the survivor rules adjusted based on the number of the total number rules. For example, in 2006 for the Developed market, the surviving rules are 290 ( $0.0137 \times 21195$ ) and their standard deviation is 598 ( $0.0282 \times 21195$ ). The average is estimated from the twelve portfolios whose OOS is on 2006. The first portfolio's IS from 01/01/2004-31/12/2005 and the remaining eleven are calculated by rolling-forward the IS by one month.

**Table C.5: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year)**

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	16.19% (2.04)	16.91% (1.97)	23.67% (1.92)	50.74% (2.18)	16.38% (2.36)	14.66% (2.13)	12.49% (2.27)	15.27% (2.28)	9.74% (2.31)	7.01% (3.14)	<b>18.31%</b> <b>(2.26)</b>
US	11.78% (1.35)	14.38% (1.56)	23.21% (1.43)	48.31% (1.65)	16.47% (1.86)	15.01% (2.21)	12.84% (2.20)	14.94% (2.03)	12.91% (1.84)	5.48% (2.33)	<b>17.53%</b> <b>(1.85)</b>
UK	10.71% (1.31)	15.37% (1.52)	18.59% (0.95)	49.79% (1.41)	22.49% (1.86)	19.64% (1.97)	7.26% (2.42)	12.81% (2.04)	11.88% (2.58)	11.17% (2.85)	<b>17.97%</b> <b>(1.89)</b>
Japan	23.33% (1.36)	13.65% (0.93)	13.77% (1.04)	34.03% (1.39)	12.38% (1.31)	15.59% (1.68)	7.84% (2.21)	16.86% (1.90)	5.81% (2.35)	8.31% (2.19)	<b>15.16%</b> <b>(1.64)</b>
<b>Emerging</b>	29.38% (2.6)	33.62% (2.51)	41.64% (2.19)	69.6% (2.51)	20.41% (2.54)	18.18% (2.35)	21.84% (2.49)	15.57% (2.59)	9.46% (2.46)	13.92% (2.48)	<b>27.36%</b> <b>(2.47)</b>
Russia	39.42% (1.26)	26.95% (1.19)	28.83% (0.95)	99.79% (1.78)	27.57% (1.43)	28.66% (2.12)	31.19% (2.26)	23.39% (2.02)	17.13% (2.23)	47.18% (2.18)	<b>37.01%</b> <b>(1.74)</b>
China	34.17% (2.22)	41.65% (2.32)	55.04% (2.10)	69.05% (1.90)	19.00% (1.86)	10.95% (1.98)	13.55% (2.70)	19.17% (2.07)	18.40% (2.34)	23.90% (1.97)	<b>30.49%</b> <b>(2.14)</b>
Brazil	28.63% (1.42)	31.85% (1.22)	40.21% (1.22)	68.74% (1.51)	24.13% (1.88)	12.54% (1.76)	29.99% (2.48)	17.52% (2.17)	21.09% (2.48)	32.27% (1.69)	<b>30.7%</b> <b>(1.78)</b>
<b>Frontier</b>	24.2% (2.69)	25.08% (2.63)	28.36% (2.27)	55.37% (2.87)	21.1% (3.21)	16.34% (2.53)	13.5% (2.30)	12.6% (2.82)	10.67% (2.85)	17.95% (2.46)	<b>22.52%</b> <b>(2.66)</b>
Estonia	18.92% (1.87)	38.79% (2.27)	44.25% (2.25)	75.54% (2.36)	38.34% (2.14)	33.91% (2.17)	15.69% (2.35)	19.76% (1.93)	11.87% (1.63)	15.01% (1.89)	<b>31.21%</b> <b>(2.09)</b>
Morocco	36.42% (3.04)	39.55% (2.48)	32.95% (2.13)	37.99% (2.16)	12.64% (2.35)	8.85% (2.31)	18.02% (1.64)	6.82% (2.35)	8.32% (2.04)	9.68% (1.31)	<b>21.12%</b> <b>(2.18)</b>
Jordan	41.04% (2.21)	29.06% (1.92)	32.41% (2.15)	46.15% (2.02)	16.23% (2.40)	17.99% (2.28)	12.73% (2.21)	8.33% (1.99)	10.38% (1.84)	5.46% (2.54)	<b>21.98%</b> <b>(2.16)</b>
<b>Average</b>	<b>26.18%</b> <b>(1.95)</b>	<b>27.24%</b> <b>(1.88)</b>	<b>31.91%</b> <b>(1.72)</b>	<b>58.76%</b> <b>(1.98)</b>	<b>20.60%</b> <b>(2.10)</b>	<b>17.69%</b> <b>(2.12)</b>	<b>16.41%</b> <b>(2.30)</b>	<b>15.25%</b> <b>(2.18)</b>	<b>12.31%</b> <b>(2.25)</b>	<b>16.45%</b> <b>(2.25)</b>	<b>24.28%</b> <b>(2.07)</b>

**Note:** The table reports the average IS annualized returns and Sharpe ratios of twelve portfolios for one year of IS after transaction costs (rolling-forward by one month). For example, the 16.19% annualized return of the Developed markets (2006) is calculated as the average IS annualized return of twelve portfolios. The first portfolio's IS return is calculated over the period of 01/01/2005-31/12/2005. The remaining eleven are calculated by rolling-forward the IS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table C.6 Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 1 Month)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	11.09% (1.27)	-15.38% (-1.66)	3.43% (0.16)	3.68% (0.25)	-8.94% (-1.96)	-13.85% (-1.67)	-2.57% (-0.87)	6.76% (1.09)	-6.95% (-1.77)	-3.07% (-1.15)	<b>-2.58%</b> <b>(-0.63)</b>
US	8.51% (0.96)	-13.79% (-1.11)	5.5% (0.21)	0.28% (0.02)	-9.76% (-1.47)	-12.72% (-1.55)	-1.17% (-0.35)	11.66% (1.55)	0.51% (0.08)	-0.56% (-0.3)	<b>-1.15%</b> <b>(-0.2)</b>
UK	16.38% (1.49)	1.7% (0.17)	7.01% (0.31)	19.95% (1.1)	-13.86% (-2.08)	-11.93% (-1.09)	-2.91% (-2.43)	-2.07% (-0.36)	-5.19% (-1.27)	-4.14% (-0.92)	<b>0.49%</b> <b>(-0.51)</b>
Japan	-10.2% (-0.63)	-4.73% (-0.77)	19.12% (0.67)	2.08% (0.09)	-7.74% (-1.41)	-13.7% (-1.64)	-2.77% (-1.62)	-2.29% (-0.19)	-6.09% (-3.11)	-4.72% (-0.91)	<b>-3.11%</b> <b>(-0.95)</b>
<b>Emerging</b>	3.72% (0.31)	7.74% (0.44)	23.23% (0.64)	0.81% (0.05)	-7.53% (-1.25)	-14.17% (-2.25)	-8.37% (-1.86)	-3.3% (-0.95)	-9.02% (-2.94)	-4.63% (-0.79)	<b>-1.15%</b> <b>(-0.86)</b>
Russia	10.33% (0.43)	-8.26% (-0.9)	49.31% (0.83)	9.41% (0.36)	-10.31% (-1.04)	-27.52% (-2.51)	-3.66% (-0.51)	-5.97% (-0.9)	8.68% (0.43)	-15.74% (-1.27)	<b>0.63%</b> <b>(-0.51)</b>
China	58.42% (2.72)	13.24% (0.57)	-0.35% (-0.01)	-13.12% (-0.69)	-10.18% (-1.41)	-19.89% (-2.41)	-2.99% (-1.57)	-6.18% (-0.93)	-8.73% (-1.12)	11.91% (0.77)	<b>2.21%</b> <b>(-0.41)</b>
Brazil	-10.06% (-0.43)	17.71% (0.67)	84.71% (1.47)	21.25% (0.98)	-12.27% (-1.62)	-18.28% (-2.17)	-14.19% (-2.37)	-3.53% (-0.63)	-4.38% (-0.36)	9.01% (0.4)	<b>7%</b> <b>(-0.41)</b>
<b>Frontier</b>	-12.23% (-1.35)	15.77% (1.53)	33.12% (1.56)	11.75% (1.15)	4.54% (1.1)	-2.12% (-0.36)	-7.06% (-2.01)	-4.42% (-1.16)	0.68% (0.13)	9.22% (1.18)	<b>4.93%</b> <b>(0.18)</b>
Estonia	-18.25% (-1.59)	-8.47% (-0.55)	68.02% (1.5)	12.74% (0.52)	1.8% (0.13)	-6.9% (-0.42)	1.27% (0.2)	-3.46% (-0.43)	6.46% (0.67)	-14.41% (-2.55)	<b>3.88%</b> <b>(-0.25)</b>
Morocco	19.55% (1.26)	7.48% (0.6)	32.43% (1.41)	-4.1% (-0.34)	-5.82% (-1.52)	-6.47% (-1.65)	2.66% (0.22)	-6.1% (-2.37)	-1.88% (-1.21)	-3.56% (-0.56)	<b>3.42%</b> <b>(-0.42)</b>
Jordan	-8.63% (-0.57)	2.5% (0.27)	39.35% (1.31)	-7.1% (-0.6)	-5.33% (-1.44)	-4.97% (-0.72)	-5.6% (-1.27)	-5.18% (-1.26)	-4.02% (-0.94)	-2.6% (-1.24)	<b>-0.16%</b> <b>(-0.64)</b>
<b>Average</b>	<b>5.72%</b> <b>(0.32)</b>	<b>1.29%</b> <b>(-0.06)</b>	<b>30.41%</b> <b>(0.84)</b>	<b>4.8%</b> <b>(0.24)</b>	<b>-7.12%</b> <b>(-1.16)</b>	<b>-12.71%</b> <b>(-1.54)</b>	<b>-3.95%</b> <b>(-1.2)</b>	<b>-2.01%</b> <b>(-0.55)</b>	<b>-2.49%</b> <b>(-0.95)</b>	<b>-1.94%</b> <b>(-0.61)</b>	<b>1.20%</b> <b>(-0.47)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of twelve portfolios for one year of IS and one month of OOS after transaction costs (rolling-forward by one month). For example, the 11.09% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over January 2006 using as IS the period 01/01/2005-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table C.7: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 3 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	-2.28% (-0.30)	-13.59% (-1.43)	11.16% (0.57)	-5.01% (-0.38)	-8.71% (-1.93)	-11.44% (-1.39)	-1.66% (-0.62)	3.71% (0.66)	-1.28% (-0.36)	-3.35% (-1.41)	<b>-3.25%</b> <b>(-0.66)</b>
US	3.83% (0.48)	-13.72% (-1.20)	19.13% (0.75)	-7.39% (-0.47)	-7.10% (-1.11)	-11.95% (-1.56)	0.11% (0.04)	8.93% (1.23)	2.91% (0.44)	-1.35% (-0.57)	<b>-0.66%</b> <b>(-0.20)</b>
UK	16.96% (1.51)	-5.58% (-0.53)	14.57% (0.52)	9.43% (0.66)	-8.81% (-1.28)	-16.99% (-1.55)	-2.29% (-1.84)	1.33% (0.25)	-4.49% (-1.10)	-1.92% (-0.44)	<b>0.22%</b> <b>(-0.38)</b>
Japan	-18.53% (-1.12)	-0.83% (-0.12)	24.25% (0.84)	-0.44% (-0.02)	-5.29% (-1.05)	-14.05% (-2.07)	-2.44% (-1.20)	-4.51% (-0.44)	-3.76% (-2.28)	-5.09% (-0.96)	<b>-3.07%</b> <b>(-0.84)</b>
<b>Emerging</b>	-6.89% (-0.60)	-0.68% (-0.04)	8.65% (0.30)	3.29% (0.26)	-5.16% (-0.91)	-5.96% (-0.91)	-5.4% (-1.24)	-3.66% (-0.95)	-6.26% (-2.31)	-7.74% (-1.39)	<b>-2.98%</b> <b>(-0.78)</b>
Russia	11.32% (0.50)	-19.86% (-1.82)	50.97% (0.92)	0.85% (0.04)	-8.58% (-0.92)	-27.03% (-2.69)	-1.60% (-0.24)	-6.99% (-1.05)	1.75% (0.10)	-14.57% (-1.25)	<b>-1.37%</b> <b>(-0.64)</b>
China	31.65% (1.52)	13.12% (0.55)	-12.97% (-0.51)	-8.78% (-0.52)	-8.99% (-1.42)	-8.86% (-1.30)	-3.07% (-1.73)	-7.99% (-1.26)	-7.77% (-1.10)	5.67% (0.40)	<b>-0.80%</b> <b>(-0.54)</b>
Brazil	-14.85% (-0.69)	24.47% (0.75)	52.01% (1.21)	11.24% (0.55)	-7.63% (-0.97)	-9.45% (-1.21)	-10.83% (-1.98)	-0.58% (-0.09)	-4.63% (-0.43)	-1.16% (-0.06)	<b>3.86%</b> <b>(-0.29)</b>
<b>Frontier</b>	-16.7% (-2.16)	7.56% (0.68)	29.16% (1.59)	-2.46% (-0.32)	3.26% (0.80)	-8.84% (-1.66)	-4.11% (-1.10)	-4.45% (-1.40)	-5.22% (-1.28)	6.75% (0.84)	<b>0.5%</b> <b>(-0.40)</b>
Estonia	-16.4% (-1.38)	-10.4% (-0.77)	65.84% (1.54)	5.71% (0.25)	-2.00% (-0.17)	-12.07% (-0.82)	4.74% (0.73)	-3.48% (-0.46)	2.14% (0.24)	-9.68% (-2.18)	<b>2.44%</b> <b>(-0.30)</b>
Morocco	6.81% (0.50)	14.73% (1.12)	11.77% (0.62)	-7.6% (-0.85)	-6.11% (-1.80)	-6.56% (-1.81)	-2.05% (-0.19)	-4.1% (-2.16)	-2.41% (-1.60)	-2.01% (-0.33)	<b>0.25%</b> <b>(-0.65)</b>
Jordan	3.76% (0.27)	5.43% (0.59)	22.25% (0.88)	-11.33% (-1.13)	-5.80% (-1.74)	-1.03% (-0.15)	-7.32% (-2.06)	-6.12% (-1.45)	-3.78% (-1.21)	-2.23% (-1.20)	<b>-0.62%</b> <b>(-0.72)</b>
<b>Average</b>	<b>-0.11%</b> <b>(-0.12)</b>	<b>0.05%</b> <b>(-0.18)</b>	<b>24.73%</b> <b>(0.77)</b>	<b>-1.04%</b> <b>(-0.16)</b>	<b>-5.91%</b> <b>(-1.04)</b>	<b>-11.19%</b> <b>(-1.43)</b>	<b>-2.99%</b> <b>(-0.95)</b>	<b>-2.33%</b> <b>(-0.59)</b>	<b>-2.73%</b> <b>(-0.91)</b>	<b>-3.06%</b> <b>(-0.71)</b>	<b>-0.46%</b> <b>(-0.53)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of four portfolios for IS of one year and OOS of three months after transaction costs (rolling-forward by one month). For example, the -2.28% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over the period 01/01/2006-31/03/2006 using as IS the period 01/01/2005-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS and the OOS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table C.8: Annualized Returns and Sharpe Ratios after Transaction Costs (IS 1 Year and OOS 6 Months)

Market	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
<b>Developed</b>	-5.63% (-0.78)	-8.98% (-0.86)	5.18% (0.31)	-2.7% (-0.25)	-5.56% (-1.30)	-8.33% (-1.14)	-0.57% (-0.23)	3.08% (0.57)	-1.39% (-0.41)	-2.00% (-0.76)	<b>-2.69%</b> <b>(-0.49)</b>
US	2.85% (0.36)	-13.17% (-1.06)	11.5% (0.53)	-3.28% (-0.25)	-4.6% (-0.77)	-7.73% (-1.10)	1.19% (0.38)	6.1% (0.89)	2.69% (0.39)	-2.28% (-0.88)	<b>-0.67%</b> <b>(-0.15)</b>
UK	17.37% (1.50)	-11.36% (-1.00)	9.35% (0.45)	7.37% (0.53)	-5.19% (-0.73)	-11.88% (-1.14)	-1.75% (-1.53)	1.02% (0.21)	-4.33% (-1.14)	-3.64% (-0.81)	<b>-0.30%</b> <b>(-0.37)</b>
Japan	-19.27% (-1.15)	-5.64% (-0.77)	18.15% (0.57)	-4.71% (-0.29)	-4.35% (-0.91)	-11.51% (-1.79)	-1.66% (-0.53)	-7.96% (-0.89)	-2.73% (-1.72)	-4.56% (-0.83)	<b>-4.42%</b> <b>(-0.83)</b>
<b>Emerging</b>	-10.01% (-0.94)	0.18% (0.01)	15.81% (0.55)	5.71% (0.47)	-2.45% (-0.45)	-8.8% (-1.30)	-2.79% (-0.59)	-6.14% (-1.48)	-5.29% (-2.08)	-4.36% (-0.73)	<b>-1.81%</b> <b>(-0.65)</b>
Russia	-1.91% (-0.11)	-22.93% (-2.14)	82.47% (1.20)	-5.17% (-0.25)	-6.1% (-0.74)	-17.8% (-1.59)	-2.53% (-0.41)	-10.39% (-1.56)	0.65% (0.04)	-8.66% (-0.71)	<b>0.76%</b> <b>(-0.63)</b>
China	26.65% (1.37)	11.83% (0.52)	-4.14% (-0.15)	-5.06% (-0.32)	-4.39% (-0.74)	-8.11% (-1.30)	-2.50% (-1.57)	-10.32% (-1.58)	0.65% (0.08)	0.00% (0.00)	<b>0.46%</b> <b>(-0.37)</b>
Brazil	-20.72% (-0.96)	31.88% (0.83)	24.5% (0.80)	8.45% (0.44)	-3.76% (-0.48)	-9.81% (-1.45)	-7.86% (-1.47)	-0.85% (-0.13)	-11.6% (-1.07)	-10.01% (-0.54)	<b>0.02%</b> <b>(-0.4)</b>
<b>Frontier</b>	-13.27% (-1.96)	3.57% (0.34)	22.57% (1.34)	-3.1% (-0.48)	2.78% (0.68)	-7.36% (-1.51)	-1.9% (-0.52)	-2.31% (-0.68)	-6.02% (-1.69)	7.67% (0.95)	<b>0.26%</b> <b>(-0.35)</b>
Estonia	-14.15% (-1.14)	-5.34% (-0.42)	37.53% (1.05)	8.79% (0.40)	-0.14% (-0.01)	-17.09% (-1.28)	2.31% (0.38)	-4.37% (-0.68)	-4.17% (-0.56)	-8.15% (-1.61)	<b>-0.48%</b> <b>(-0.39)</b>
Morocco	1.15% (0.09)	14.15% (1.06)	-0.90% (-0.06)	-7.89% (-0.97)	-5.22% (-1.49)	-4.44% (-1.19)	-2.55% (-0.26)	-1.92% (-1.01)	-2.92% (-1.74)	-5.03% (-0.82)	<b>-1.56%</b> <b>(-0.64)</b>
Jordan	4.05% (0.31)	3.48% (0.38)	19.08% (0.84)	-11.84% (-1.27)	-4.89% (-1.33)	-1.70% (-0.26)	-7.22% (-2.06)	-5.10% (-1.05)	-3.28% (-1.27)	-2.66% (-1.41)	<b>-1.01%</b> <b>(-0.71)</b>
<b>Average</b>	<b>-2.74%</b> <b>(-0.28)</b>	<b>-0.19%</b> <b>(-0.26)</b>	<b>20.09%</b> <b>(0.62)</b>	<b>-1.12%</b> <b>(-0.19)</b>	<b>-3.65%</b> <b>(-0.69)</b>	<b>-9.55%</b> <b>(-1.26)</b>	<b>-2.32%</b> <b>(-0.70)</b>	<b>-3.26%</b> <b>(-0.62)</b>	<b>-3.14%</b> <b>(-0.93)</b>	<b>-3.64%</b> <b>(-0.68)</b>	<b>-0.95%</b> <b>(-0.50)</b>

**Note:** The table reports the average OOS annualized returns and Sharpe ratios of four portfolios for IS of one year and OOS of six months after transaction costs (rolling-forward by one month). For example, the -5.63% annualized return of the Developed markets (2006) is calculated as the average OOS annualized return of twelve portfolios. The first portfolio's OOS return is calculated over the period 01/01/2006-31/06/2006 using as IS the period 01/01/2005-31/12/2005. The remaining eleven OOS returns are calculated by rolling-forward the IS and the OOS by one month. The same logic applies to the Sharpe ratios. The last column and row presents the average performance per market across all years and per year respectively.

Table D.1: Specification of the Volatility Forecasting Pool

Class	Count	Family	Error Distribution	Mean	Variance
1. ARCH	48	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
2. GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
3. IGARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
4. Taylor/ Schwert	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
5. A-GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
6. NA-GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
7. TGARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
8. GJR-GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
9. log-GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
10. EGARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
11. NGARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
12. APARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
13. FI-GARCH	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
14. GARCH-MA	96	GARCH	Gaussian, t, skewed-t & GED	Zero, unconditional and conditional	RV & ARV
15. SV	36	SV	Gaussian, t & skewed-t	Zero, unconditional and conditional	RV & ARV
16. SV-MA	36	SV	Gaussian, t & skewed-t	Zero, unconditional and conditional	RV & ARV
17. SV-L	72	SV	Gaussian, t & skewed-t	Zero, unconditional and conditional	RV & ARV
18. RM	60	EWMA	-	Zero, unconditional and conditional	RV & ARV
19. HAR	6	HAR	-	Zero, unconditional and conditional	RV & ARV
20. log-HAR	6	HAR	-	Zero, unconditional and conditional	RV & ARV
<b>Total</b>	<b>1512</b>				



Table D.2: Variation in Number of True Discoveries for  $MAE_1$ 

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	9	573	1013	489	893	595.4
	GARCH (1,1)	27	520	1	512	96	231.2
	PRC 90	5	16	1	81	96	39.8
GBP/USD	ARCH (1)	450	1	751	178	132	302.4
	GARCH (1,1)	102	1	2	113	132	70
	PRC 90	96	1	1	97	132	65.4
USD/JPY	ARCH (1)	457	1	1	1162	346	393.4
	GARCH (1,1)	1	42	1	701	296	208.2
	PRC 90	1	1	1	1	46	10
DJIA	ARCH (1)	1	39	137	3	39	43.8
	GARCH (1,1)	1	116	268	3	34	84.4
	PRC 90	113	93	106	18	48	75.6
FTSE 100	ARCH (1)	273	1	1	1	1	55.4
	GARCH (1,1)	274	48	276	1	19	123.6
	PRC 90	102	2	82	1	33	44
XAU/USD	ARCH (1)	433	1	716	819	3	394.4
	GARCH (1,1)	526	1	663	43	102	267
	PRC 90	5	2	1	2	1	2.2

**Note:** The table presents the size of the rejection set for different markets based on the  $MAE_1$  benchmark. The  $DFDR^+$  always reject the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table D.3: Variation in Number of True Discoveries for  $MAE_2$ 

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	9	560	1018	493	894	594.8
	GARCH (1,1)	28	521	1	513	96	231.8
	PRC 90	6	21	1	75	96	39.8
GBP/USD	ARCH (1)	477	1	760	177	132	309.4
	GARCH (1,1)	102	1	1	113	132	69.8
	PRC 90	96	1	1	97	132	65.4
USD/JPY	ARCH (1)	375	1	1	1164	347	377.6
	GARCH (1,1)	1	7	1	649	296	190.8
	PRC 90	1	1	1	1	50	10.8
DJIA	ARCH (1)	1	39	166	3	38	49.4
	GARCH (1,1)	1	126	302	3	38	94
	PRC 90	113	91	107	18	48	75.4
FTSE 100	ARCH (1)	272	8	1	1	1	56.6
	GARCH (1,1)	277	48	335	1	19	136
	PRC 90	104	1	77	1	32	43
XAU/USD	ARCH (1)	405	1	736	780	1	384.6
	GARCH (1,1)	458	3	687	41	7	239.2
	PRC 90	5	1	1	1	1	1.8

**Note:** The table presents the size of the rejection set for different markets based on the  $MAE_2$  benchmark. The  $DFDR^+$  always reject the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table D.4: Variation in Number of True Discoveries for  $MSE_2$ 

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	5	482	984	463	763	539.4
	GARCH (1,1)	13	447	1	481	96	207.6
	PRC 90	1	25	1	46	96	33.8
GBP/USD	ARCH (1)	290	1	720	174	87	254.4
	GARCH (1,1)	97	4	2	112	101	63.2
	PRC 90	96	1	1	93	91	56.4
USD/JPY	ARCH (1)	275	1	1	996	336	321.8
	GARCH (1,1)	1	364	1	545	282	238.6
	PRC 90	5	3	1	1	10	4
DJIA	ARCH (1)	1	34	114	52	38	47.8
	GARCH (1,1)	3	122	358	53	39	115
	PRC 90	84	102	109	59	40	78.8
FTSE 100	ARCH (1)	189	19	1	1	3	42.6
	GARCH (1,1)	199	48	440	1	17	141
	PRC 90	87	36	127	1	65	63.2
XAU/USD	ARCH (1)	327	1	502	699	2	306.2
	GARCH (1,1)	426	3	475	23	158	217
	PRC 90	56	3	1	1	2	12.6

**Note:** The table presents the size of the rejection set for different markets based on the  $MSE_2$  benchmark. The  $DFDR^+$  always reject the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table D.5: Variation in Number of True Discoveries for  $R^2\text{LOG}$ 

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	158	810	1101	570	1098	747.4
	GARCH (1,1)	173	755	178	581	217	380.8
	PRC 90	118	113	65	133	75	100.8
GBP/USD	ARCH (1)	699	169	835	484	143	466
	GARCH (1,1)	195	341	364	320	143	272.6
	PRC 90	100	97	22	102	6	65.4
USD/JPY	ARCH (1)	804	68	54	1344	440	542
	GARCH (1,1)	78	314	54	962	413	364.2
	PRC 90	101	68	52	64	66	70.2
DJIA	ARCH (1)	133	352	621	247	79	286.4
	GARCH (1,1)	76	374	619	230	80	275.8
	PRC 90	122	82	116	104	78	100.4
FTSE 100	ARCH (1)	960	244	39	337	90	334
	GARCH (1,1)	958	297	50	221	108	326.8
	PRC 90	85	115	39	110	70	83.8
XAU/USD	ARCH (1)	932	102	821	1098	518	694.2
	GARCH (1,1)	906	140	773	573	678	614
	PRC 90	75	84	101	112	85	91.4

**Note:** The table presents the size of the rejection set for different markets based on the  $R^2\text{LOG}$  benchmark. The  $\text{DFDR}^+$  always reject the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table D.6: Variation in Number of True Discoveries for QLIKE

Asset	Benchmark	Study Period					Average
		2013	2014	2015	2016	2017	
EUR/USD	ARCH (1)	2	498	959	469	528	491.2
	GARCH (1,1)	10	489	1	491	96	217.4
	PRC 90	1	12	1	73	96	36.6
GBP/USD	ARCH (1)	212	1	517	178	102	202
	GARCH (1,1)	96	181	1	109	114	100.2
	PRC 90	96	1	1	97	110	61
USD/JPY	ARCH (1)	536	1	1	886	326	350
	GARCH (1,1)	1	485	2	489	267	248.8
	PRC 90	4	1	1	1	10	3.4
DJIA	ARCH (1)	1	34	92	18	31	35.2
	GARCH (1,1)	1	111	247	18	38	83
	PRC 90	102	103	126	31	39	80.2
FTSE 100	ARCH (1)	174	24	1	1	5	41
	GARCH (1,1)	185	47	253	1	22	101.6
	PRC 90	92	42	89	80	64	73.4
XAU/USD	ARCH (1)	326	1	418	497	2	248.8
	GARCH (1,1)	447	3	411	30	230	224.2
	PRC 90	56	1	1	1	2	12.2

Note: The table presents the size of the rejection set for different markets based on the QLIKE benchmark. The DFDR<sup>+</sup> always reject the lowest  $p$ -value before any further computations. Therefore, the cases with 1 discovery can be interpreted as no discoveries at all. The ARCH (1) and GARCH (1,1) benchmarks are zero mean, with Gaussian distribution and RV as the conditional variance specification. PRC 90 corresponds to the 90<sup>th</sup> percentile of the entire volatility pool.

Table D.7: Innovation Distribution for  $MAE_1$ 

Asset	Benchmark	Gaussian	t	Skewed t	GED	No dist.
EUR/USD	ARCH (1)	40.70%	40.27%	40.32%	<b>42.84%</b>	7.50%
	GARCH (1,1)	<b>16.24%</b>	15.70%	15.91%	15.56%	3.89%
	PRC 90	2.96%	2.74%	<b>3.12%</b>	2.16%	0.00%
GBP/USD	ARCH	20.32%	20.32%	20.32%	<b>22.59%</b>	3.33%
	GARCH (1,1)	4.57%	4.62%	<b>5.32%</b>	4.57%	1.67%
	PRC 90	4.52%	4.52%	<b>4.68%</b>	4.44%	0.00%
USD/JPY	ARCH	28.92%	26.61%	20.11%	<b>31.79%</b>	12.50%
	GARCH (1,1)	14.89%	14.19%	10.97%	<b>16.91%</b>	6.11%
	PRC 90	1.02%	0.48%	0.11%	<b>1.23%</b>	0.00%
DJIA	ARCH	3.01%	<b>3.28%</b>	2.58%	3.15%	0.83%
	GARCH (1,1)	5.97%	<b>6.77%</b>	5.00%	5.49%	0.83%
	PRC 90	5.54%	<b>6.08%</b>	5.22%	4.01%	0.00%
FTSE 100	ARCH	3.87%	<b>4.03%</b>	3.82%	3.58%	0.28%
	GARCH (1,1)	8.23%	8.71%	<b>9.78%</b>	7.04%	1.94%
	PRC 90	2.96%	3.06%	<b>3.49%</b>	2.65%	0.00%
XAU/USD	ARCH	23.28%	28.33%	<b>31.51%</b>	25.56%	3.33%
	GARCH (1,1)	15.32%	18.23%	<b>23.33%</b>	16.79%	1.39%
	PRC 90	0.05%	0.00%	<b>0.43%</b>	0.12%	0.00%
Average		11.24%	11.55%	11.45%	<b>11.69%</b>	2.42%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 40.7% means that on average 151 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_1$ . The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.

Table D.8: Innovation Distribution for  $MAE_2$ 

Asset	Benchmark	Gaussian	t	Skewed t	GED	No Dist.
EUR/USD	ARCH (1)	40.59%	40.22%	40.22%	<b>42.84%</b>	8.06%
	GARCH (1,1)	<b>16.18%</b>	15.75%	15.97%	15.68%	3.89%
	PRC 90	2.90%	2.69%	<b>3.12%</b>	2.28%	0.00%
GBP/USD	ARCH (1)	20.97%	20.70%	20.91%	<b>22.78%</b>	3.89%
	GARCH (1,1)	4.57%	4.62%	<b>5.27%</b>	4.57%	1.67%
	PRC 90	4.52%	4.52%	<b>4.68%</b>	4.44%	0.00%
USD/JPY	ARCH (1)	27.80%	25.32%	19.62%	<b>30.25%</b>	12.50%
	GARCH (1,1)	13.66%	12.80%	9.78%	<b>15.93%</b>	6.11%
	PRC 90	1.02%	0.59%	0.16%	<b>1.30%</b>	0.00%
DJIA	ARCH (1)	3.49%	<b>3.66%</b>	2.85%	3.58%	0.83%
	GARCH (1,1)	6.77%	<b>7.69%</b>	5.38%	6.05%	0.83%
	PRC 90	5.59%	<b>6.08%</b>	5.16%	3.95%	0.00%
FTSE 100	ARCH (1)	3.92%	4.03%	<b>4.03%</b>	3.52%	0.83%
	GARCH (1,1)	9.03%	9.46%	<b>10.38%</b>	8.21%	2.78%
	PRC 90	3.01%	2.90%	<b>3.49%</b>	2.47%	0.00%
XAU/USD	ARCH (1)	23.01%	27.04%	<b>30.43%</b>	25.56%	3.33%
	GARCH (1,1)	14.68%	15.97%	<b>19.52%</b>	15.93%	1.39%
	PRC 90	0.00%	0.00%	<b>0.38%</b>	0.12%	0.00%
Average		11.21%	11.34%	11.19%	<b>11.64%</b>	2.56%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 40.59% means that on average 150 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_1$ . The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.

Table D.9: Innovation Distribution for  $MSE_2$ 

Asset	Benchmark	Gaussian	t	Skewed t	GED	No Dist.
EUR/USD	ARCH (1)	36.94%	35.97%	36.67%	<b>39.81%</b>	3.89%
	GARCH (1,1)	14.41%	13.98%	14.30%	<b>14.63%</b>	1.94%
	PRC 90	2.31%	2.15%	<b>2.74%</b>	2.16%	0.00%
GBP/USD	ARCH (1)	17.31%	17.80%	16.13%	<b>19.32%</b>	1.67%
	GARCH (1,1)	4.09%	4.25%	<b>4.78%</b>	4.38%	0.28%
	PRC 90	3.76%	3.87%	3.92%	<b>4.14%</b>	0.00%
USD/JPY	ARCH (1)	23.87%	21.61%	16.67%	<b>26.54%</b>	6.39%
	GARCH (1,1)	16.83%	16.94%	13.82%	<b>18.46%</b>	2.50%
	PRC 90	0.38%	0.11%	0.16%	<b>0.49%</b>	0.00%
DJIA	ARCH (1)	3.33%	<b>3.98%</b>	3.44%	2.22%	0.83%
	GARCH (1,1)	8.23%	<b>8.92%</b>	7.53%	6.60%	2.50%
	PRC 90	5.38%	<b>6.72%</b>	5.81%	3.58%	0.83%
FTSE 100	ARCH (1)	3.06%	<b>3.17%</b>	2.85%	2.59%	0.56%
	GARCH (1,1)	9.41%	9.95%	<b>10.86%</b>	8.09%	3.33%
	PRC 90	4.35%	<b>4.84%</b>	4.78%	3.40%	0.28%
XAU/USD	ARCH (1)	18.87%	20.38%	<b>25.16%</b>	20.19%	1.67%
	GARCH (1,1)	12.26%	14.68%	<b>20.43%</b>	12.35%	1.11%
	PRC 90	0.59%	0.70%	<b>1.77%</b>	0.37%	0.00%
<b>Average</b>		10.30%	10.56%	<b>10.66%</b>	10.52%	1.54%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 36.94% means that on average 137 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $MSE_1$ . The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.



Table D.10: Innovation Distribution for  $R^2\text{LOG}$ 

Asset	Benchmark	Gaussian	t	Skewed t	GED	No Dist.
EUR/USD	ARCH (1)	<b>53.44%</b>	50.05%	50.05%	51.36%	13.61%
	GARCH (1,1)	<b>28.28%</b>	25.05%	26.18%	24.63%	7.22%
	PRC 90	<b>8.71%</b>	7.15%	6.61%	4.94%	1.67%
GBP/USD	ARCH (1)	<b>33.76%</b>	31.61%	31.77%	30.62%	7.50%
	GARCH (1,1)	<b>20.65%</b>	19.78%	18.17%	15.93%	4.17%
	PRC 90	<b>5.32%</b>	4.30%	4.41%	4.07%	0.00%
USD/JPY	ARCH (1)	39.95%	35.54%	29.41%	<b>41.60%</b>	23.61%
	GARCH (1,1)	<b>27.31%</b>	24.41%	19.78%	26.91%	15.28%
	PRC 90	<b>6.45%</b>	4.30%	3.17%	5.68%	0.00%
DJIA	ARCH (1)	20.32%	18.82%	17.58%	<b>20.56%</b>	12.22%
	GARCH (1,1)	19.25%	17.80%	17.42%	<b>19.75%</b>	12.78%
	PRC 90	<b>7.80%</b>	6.56%	5.48%	7.53%	3.06%
FTSE 100	ARCH (1)	21.88%	21.18%	22.69%	<b>23.77%</b>	17.22%
	GARCH (1,1)	21.88%	20.32%	22.04%	<b>23.33%</b>	16.94%
	PRC 90	5.38%	5.54%	5.05%	6.05%	<b>6.67%</b>
XAU/USD	ARCH (1)	45.81%	45.65%	<b>49.19%</b>	48.64%	18.61%
	GARCH (1,1)	39.25%	41.34%	<b>47.10%</b>	39.20%	16.67%
	PRC 90	<b>7.42%</b>	6.45%	6.02%	5.12%	1.11%
<b>Average</b>		<b>22.94%</b>	21.44%	21.23%	22.21%	9.91%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 53.44% means that on average 199 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is  $\text{MSE}_1$ . The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.

Table D.11: Innovation Distribution for QLIKE

Asset	Benchmark	Gaussian	t	Skewed t	GED	No Dist.
EUR/USD	ARCH (1)	34.09%	32.69%	<b>34.89%</b>	34.26%	2.78%
	GARCH (1,1)	15.38%	15.05%	<b>15.65%</b>	13.95%	1.11%
	PRC 90	2.69%	2.42%	<b>2.96%</b>	2.04%	0.00%
GBP/USD	ARCH (1)	14.09%	13.98%	13.33%	<b>14.44%</b>	1.67%
	GARCH (1,1)	6.88%	7.31%	<b>7.80%</b>	5.68%	0.00%
	PRC 90	4.14%	<b>4.46%</b>	4.41%	3.89%	0.00%
USD/JPY	ARCH (1)	26.02%	23.98%	18.06%	<b>28.95%</b>	4.17%
	GARCH (1,1)	17.63%	17.58%	14.95%	<b>19.07%</b>	0.56%
	PRC 90	0.38%	0.05%	0.05%	<b>0.49%</b>	0.00%
DJIA	ARCH (1)	2.74%	<b>3.06%</b>	2.26%	1.60%	0.00%
	GARCH (1,1)	6.08%	<b>6.24%</b>	5.38%	5.06%	1.11%
	PRC 90	5.91%	<b>6.61%</b>	5.81%	3.70%	0.00%
FTSE 100	ARCH (1)	2.85%	<b>3.17%</b>	2.80%	2.47%	0.28%
	GARCH (1,1)	6.88%	7.31%	<b>8.01%</b>	5.43%	1.94%
	PRC 90	5.27%	<b>5.54%</b>	5.48%	3.95%	0.00%
XAU/USD	ARCH (1)	14.84%	16.67%	<b>23.44%</b>	13.58%	0.56%
	GARCH (1,1)	12.42%	15.48%	<b>21.24%</b>	12.65%	0.56%
	PRC 90	0.54%	0.70%	<b>1.67%</b>	0.43%	0.00%
<b>Average</b>		9.93%	10.13%	<b>10.45%</b>	9.54%	0.82%

**Note:** The table presents the average proportion of models with each distribution able to beat the three benchmarks. For example, the first value 34.09% means that on average 127 models out of 372 Gaussian models outperformed the ARCH (1) benchmark for the EUR/USD. The ARCH (1) and GARCH (1,1) models use zero mean, Gaussian distribution and RV specifications. PRC 90 stands for the benchmark based on the 90<sup>th</sup> percentile of the entire volatility pool. The performance scale is MSE<sub>1</sub>. The 'No dist.' column corresponds to models without any distribution and the value in bold is the maximum in each row.

Table D.12: Classes Survival Analysis for  $MAE_1$ 

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	11.39%	1.11%	9.44%	17.64%	<b>22.50%</b>	17.50%	13.26%
GARCH	20.63%	3.19%	15.00%	3.61%	10.63%	16.32%	11.56%
IGARCH	<b>48.75%</b>	<b>66.88%</b>	<b>31.18%</b>	7.36%	3.13%	<b>34.93%</b>	<b>32.04%</b>
Taylor/Schwert	21.25%	13.47%	17.99%	9.93%	6.25%	23.61%	15.42%
A-GARCH	20.97%	3.19%	16.46%	0.56%	0.21%	14.51%	9.32%
NA-GARCH	21.25%	3.19%	16.11%	0.56%	0.35%	14.86%	9.39%
TGARCH	30.14%	16.25%	17.15%	0.28%	0.14%	15.28%	13.21%
GJR-GARCH	22.64%	3.19%	12.92%	0.00%	0.00%	10.56%	8.22%
log-GARCH	0.35%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%
EGARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NGARCH	22.01%	13.75%	20.00%	1.53%	1.81%	21.67%	13.46%
APARCH	22.36%	13.54%	20.76%	0.56%	0.28%	16.81%	12.38%
FI-GARCH	13.26%	1.81%	10.69%	12.29%	15.83%	19.79%	12.28%
GARCH-MA	20.14%	3.40%	16.39%	4.44%	11.81%	19.65%	12.64%
SV	11.11%	1.11%	11.11%	<b>26.67%</b>	20.56%	15.56%	14.35%
SV-MA	11.11%	1.11%	11.11%	26.11%	20.00%	15.56%	14.17%
SV-L	26.94%	10.00%	0.00%	0.83%	0.00%	1.11%	6.48%
RM	4.56%	2.00%	7.44%	0.67%	0.89%	1.89%	2.91%
HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
log-HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 11.39% means that on average 5 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is  $MAE_1$  and the value in bold shows the maximum of each column.

Table D.13: Classes Survival Analysis for  $MAE_2$ 

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	11.39%	0.97%	7.78%	18.47%	<b>22.22%</b>	15.14%	12.66%
GARCH	20.69%	3.26%	14.44%	4.58%	11.18%	15.63%	11.63%
IGARCH	<b>49.79%</b>	<b>66.81%</b>	<b>30.83%</b>	7.64%	3.06%	<b>33.33%</b>	<b>31.91%</b>
Taylor/Schwert	21.25%	13.75%	17.29%	10.35%	7.43%	22.08%	15.36%
A-GARCH	21.04%	3.26%	16.04%	0.56%	0.35%	14.10%	9.22%
NA-GARCH	21.39%	3.26%	15.07%	0.56%	0.56%	13.61%	9.07%
TGARCH	30.49%	17.15%	15.90%	0.28%	0.14%	13.47%	12.91%
GJR-GARCH	23.40%	3.19%	11.74%	0.00%	0.00%	9.10%	7.91%
log-GARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NGARCH	22.01%	13.96%	19.17%	2.71%	2.08%	21.94%	13.65%
APARCH	22.50%	13.75%	19.79%	0.56%	0.56%	16.32%	12.25%
FI-GARCH	13.33%	2.01%	10.00%	13.26%	16.11%	18.33%	12.18%
GARCH-MA	20.21%	3.75%	15.83%	5.21%	12.50%	18.47%	12.66%
SV	11.11%	1.11%	8.89%	<b>27.78%</b>	21.11%	15.00%	14.17%
SV-MA	11.11%	1.11%	8.33%	25.56%	21.11%	15.00%	13.70%
SV-L	23.61%	10.00%	0.00%	0.83%	0.00%	1.11%	5.93%
RM	4.78%	2.22%	7.44%	0.67%	1.44%	1.89%	3.07%
HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
log-HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 11.39% means that on average 5 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is  $MAE_2$  and the value in bold shows the maximum of each column.

Table D.14: Classes Survival Rate for  $MSE_2$ 

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	11.11%	1.53%	14.03%	17.08%	23.19%	17.92%	14.14%
GARCH	19.44%	3.06%	14.44%	3.82%	8.89%	15.00%	10.78%
IGARCH	<b>43.96%</b>	<b>65.14%</b>	<b>27.64%</b>	7.29%	1.94%	<b>24.44%</b>	<b>28.40%</b>
Taylor/Schwert	20.14%	11.04%	17.85%	12.64%	7.08%	20.83%	14.93%
A-GARCH	19.58%	2.99%	13.47%	0.35%	1.46%	8.06%	7.65%
NA-GARCH	20.00%	2.99%	11.32%	0.35%	1.18%	8.40%	7.37%
TGARCH	28.33%	11.11%	15.07%	0.42%	0.69%	8.68%	10.72%
GJR-GARCH	18.82%	2.85%	10.90%	0.00%	0.07%	7.08%	6.62%
log-GARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NGARCH	21.46%	12.08%	17.36%	4.10%	2.57%	17.78%	12.56%
APARCH	21.46%	11.04%	17.08%	0.42%	0.83%	12.36%	10.53%
FI-GARCH	12.64%	1.74%	14.44%	17.29%	17.22%	22.85%	14.36%
GARCH-MA	19.31%	3.13%	16.32%	5.14%	11.53%	17.78%	12.20%
SV	11.11%	1.11%	14.44%	<b>30.00%</b>	<b>26.11%</b>	17.78%	16.76%
SV-MA	11.11%	1.11%	14.44%	28.33%	26.11%	17.22%	16.39%
SV-L	14.17%	0.83%	0.00%	0.83%	0.00%	0.00%	2.64%
RM	2.33%	0.78%	3.56%	1.67%	1.67%	1.11%	1.85%
HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
log-HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 11.11% means that on average 5 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is  $MSE_2$  and the value in bold shows the maximum of each column.

Table D.15: Classes Survival Rate for  $R^2\text{LOG}$ 

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	11.53%	7.64%	12.64%	11.25%	12.64%	28.89%	14.10%
GARCH	20.76%	3.89%	18.75%	12.92%	15.14%	27.43%	16.48%
IGARCH	47.29%	59.10%	37.01%	22.57%	23.54%	<b>55.76%</b>	40.88%
Taylor/Schwert	22.78%	19.38%	22.78%	13.19%	13.40%	36.53%	21.34%
A-GARCH	21.46%	5.00%	20.69%	18.68%	20.83%	27.78%	19.07%
NA-GARCH	22.78%	5.07%	20.28%	18.54%	24.72%	29.03%	20.07%
TGARCH	31.04%	21.67%	22.43%	<b>24.44%</b>	26.53%	37.78%	27.31%
GJR-GARCH	25.69%	3.26%	20.69%	19.58%	<b>27.57%</b>	28.19%	20.83%
log-GARCH	25.97%	27.43%	20.83%	5.69%	7.22%	19.93%	17.85%
EGARCH	45.90%	24.58%	11.94%	4.44%	5.00%	12.15%	17.34%
NGARCH	22.85%	19.93%	24.03%	12.71%	13.61%	36.32%	21.57%
APARCH	25.07%	20.35%	23.68%	18.68%	20.28%	38.61%	24.44%
FI-GARCH	13.47%	3.82%	14.03%	13.75%	14.86%	27.71%	14.61%
GARCH-MA	21.11%	5.63%	20.83%	17.22%	15.14%	32.50%	18.74%
SV	11.11%	7.78%	14.44%	11.11%	12.22%	26.67%	13.89%
SV-MA	11.11%	7.78%	14.44%	11.11%	12.22%	26.67%	13.89%
SV-L	<b>81.11%</b>	<b>63.33%</b>	<b>45.56%</b>	8.89%	6.67%	43.61%	<b>41.53%</b>
RM	7.00%	4.00%	15.56%	9.44%	15.67%	13.44%	10.85%
HAR	20.00%	6.67%	0.00%	10.00%	6.67%	3.33%	7.78%
log-HAR	20.00%	6.67%	0.00%	10.00%	6.67%	3.33%	7.78%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 11.53% means that on average 6 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is  $R^2\text{LOG}$  and the value in bold shows the maximum of each column.

Table D.16: Classes Survival Rate for QLIKE

Class	EUR/USD	GBP/USD	USD/JPY	DJIA	FTSE 100	XAU/USD	Average
ARCH	10.83%	7.64%	14.17%	18.19%	24.17%	20.14%	15.86%
GARCH	16.81%	1.18%	14.51%	2.43%	7.99%	12.85%	9.29%
IGARCH	<b>41.46%</b>	<b>59.31%</b>	<b>31.18%</b>	6.32%	1.18%	22.29%	<b>26.96%</b>
Taylor/Schwert	18.61%	10.00%	17.85%	8.75%	4.10%	18.26%	12.93%
A-GARCH	16.88%	1.25%	14.17%	0.49%	0.14%	6.11%	6.50%
NA-GARCH	17.50%	1.25%	14.38%	0.49%	0.14%	6.04%	6.63%
TGARCH	23.75%	6.46%	15.90%	0.35%	0.07%	7.01%	8.92%
GJR-GARCH	14.86%	1.18%	13.82%	0.00%	0.00%	6.18%	6.01%
log-GARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NGARCH	19.65%	10.63%	18.06%	1.18%	2.15%	16.11%	11.30%
APARCH	19.79%	9.65%	19.51%	0.28%	0.14%	10.42%	9.97%
FI-GARCH	12.15%	5.63%	14.72%	12.71%	16.04%	<b>22.71%</b>	13.99%
GARCH-MA	16.94%	2.01%	15.49%	2.99%	10.42%	16.39%	10.71%
SV	11.11%	7.78%	15.56%	<b>31.11%</b>	<b>26.67%</b>	18.89%	18.52%
SV-MA	11.11%	7.78%	14.44%	30.00%	26.67%	17.78%	17.96%
SV-L	34.17%	10.00%	0.00%	0.83%	0.00%	0.00%	7.50%
RM	1.56%	0.67%	1.89%	0.44%	0.89%	0.44%	0.98%
HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
log-HAR	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

**Note:** The table presents the average proportion of each class of volatility models able to beat the three benchmarks. For example, the first value 10.83% means that on average 5 models out of 48 ARCH models outperformed the three benchmarks for the EUR/USD. The equation and the count of models for all classes are given in Table 5.1 and Table D.1 respectively. The performance scale is QLIKE and the value in bold shows the maximum of each column.